

Manifold learning and differential geometry for disease detection and risk prediction

Ashish Kumar [§]

Department of Computer Science and Engineering

Manipal University Jaipur

Jaipur 303007

Rajasthan

India

Komal Mehta Bhagat [†]

Department of Electronics & Communication Engineering

Bhagwan Parshuram Institute of Technology

Rohini 110089

Delhi

India

Pankaj Dadheech ^{*}

Department of Computer Science and Engineering

Swami Keshvanand Institute of Technology, Management & Gramothan

Jaipur 302017

Rajasthan

India

Sanwta Ram Dogiwal [‡]

Department of Information Technology

Swami Keshvanand Institute of Technology, Management & Gramothan

Jaipur 302017

Rajasthan

India

[§] E-mail: kumar.ashish@jaipur.manipal.edu

[†] E-mail: komalbhagat@bpitindia.com

^{*} E-mail: pankajdadheech777@gmail.com (Corresponding Author)

[‡] E-mail: dogiwal@gmail.com

Rakesh Sharma^{*}

Department of Computer Science and Engineering
Vivekanand Global University
Jaipur 303012
Rajasthan
India

Abstract

Healthcare datasets pose unique challenges: they are large, varied, and often contain hidden structure that standard analysis methods struggle to reveal. In this work, we propose a practical approach that combines tools from manifold learning (for example, Isomap or t-SNE) with ideas borrowed from differential geometry (such as curvature and geodesics). First, we use manifold techniques to reduce the dimensionality of complex, multi-modal inputs—ranging from wearable sensors and imaging scans to electronic health records while still preserving the data’s underlying shape. Next, by examining the resulting low-dimensional “surface” through a geometric lens, we extract features that shed light on subtle relationships among patients’ measurements, relationships that might otherwise go unnoticed. These geometric insights are then woven into a deep neural network, effectively guiding the model toward more reliable predictions of disease presence and future health risks. In our experiments, this hybrid strategy yields noticeably higher accuracy for early detection compared to conventional methods. Importantly, because we can trace predictions back to geometric properties such as how data points “bend” or “connect” on the inferred manifold clinicians gain a clearer picture of why certain patients are flagged as high risk. Beyond the technical gains, this framework is well suited for real-world healthcare settings, including resource-limited environments common in parts of India. By focusing on interpretable, scalable methods, we aim to deliver a solution that can be adopted without extensive computational resources. Looking ahead, we plan to refine the pipeline for near-real-time use enabling, for instance, community health workers to monitor groups of patients at once and to explore applications in population-level screening and preventative healthcare.

Subject Classification: 62H25.

Keywords: *Manifold learning, Differential geometry, Disease detection, Geometric deep learning, High-dimensional data, AI in Healthcare.*

1. Introduction

Additional massive and complex medical data have brought lots of challenges to both medical researchers and clinicians. Whether it’s eHealth records (EHRs), medical images, or outputs from wearable devices, data is typically high-dimensional, non-linear and heterogeneous. This implicitly requires investigating more sophisticated mathematical structures able to deal with this complex system. Manifold learning and differential

^{*} E-mail: srak.911@gmail.com

geometry have recently shown great potential for solving these problems. Manifold learning offers the possibility of reducing dimensionality by revealing low-dimensional latent spaces embedded in high-dimensional data while respecting the underlying geometry and topology of the space that the data is drawn from. On the other hand, differential geometry is a powerful mathematical tool to study the curvature, geodesics and other intrinsic phenomena in these manifolds. Combined together, they can form a better model to handle complex health datasets for improved disease detection and risk prediction [1, 2].

The objective of this paper is to propose a mathematical model that unifies manifold learning and differential geometry for disease detection and risk assessment. The framework proposed uses the geometric characteristics of medical high-dimensional data to extract and emphasize latent patterns that enhance the interpretability and robustness of AI-based predictions. This is especially important for adopted real-world healthcare scenarios, where interpretability and robustness are crucial.

2. Mathematical Foundations

The theoretical backbone of this framework lies in the principles of manifold learning and differential geometry. A manifold M is defined as a topological space that locally resembles Euclidean space $\mathbb{R}^n$. This allows high-dimensional data to be represented in a more interpretable low-dimensional space while retaining its intrinsic geometric properties. Formally, a manifold is a set of points $\{x_1, x_2, \dots, x_n\}$ equipped with a coordinate chart $\phi: M \rightarrow \mathbb{R}^n$, where ϕ is a homeomorphism. Manifold learning focuses on identifying the mapping ϕ and preserving distances between points using techniques such as:

Isomap: Preserves geodesic distances $d_g(x_i, x_j)$ by approximating the shortest paths on a graph constructed from the data. Mathematically

$$dg(x_i, x_j) = \min_{\text{path } \gamma} \int \|\dot{\gamma}(t)\| dt,$$

where $\gamma(t)$ is a path connecting x_i and x_j

t-SNE: Minimizes the Kullback-Leibler divergence between probability distributions P and Q defined on high- and low-dimensional spaces:

$$KL(P \parallel Q) = \sum_{ij} P_{ij} \log P_{ij} / Q_{ij}$$

where P_{ij} and Q_{ij} represent pairwise similarities in high- and low-dimensional spaces, respectively.

Differential geometry provides additional tools for analyzing the curvature and topology of these manifolds. The Riemannian metric tensor g_{ij} defines the local inner product on the tangent space T_pM of a point p on the manifold

$$ds^2 = \sum_{ij} g_{ij} dx^i dx^j$$

where ds represents the infinitesimal distance. The geodesic equation, given by

$$\frac{d^2 x^i}{dt^2} + \Gamma_{jk}^i \frac{dx^j}{dt} \frac{dx^k}{dt} = 0,$$

describes the shortest path between two points on the manifold, where Γ_{jk}^i are the Christoffel symbols derived from the metric tensor.

3. Manifold Learning and Differential Geometry

Most of medical data resides in high dimensional space due to the recorded information from various features using tools of diagnosis (e.g., imaging systems or wearable devices) and diagnosis platforms (e.g., genomic sequencing platforms). The inherent dimensionality of this data however is generally much smaller, since the variables often exhibit correlation, or are conditionally dependent on some underlying factors. The discovery and interpretation of these types of latent structures are unwieldy but necessary in uncovering useful patterns and predicting accurately. Here is where manifold learning and differential geometry come in the picture. Manifold learning is a family of powerful mathematical tools to discover hidden structures in high-dimensional data. This approach preserves the intrinsic geometric and topological features of the original data by determining the low-dimensional manifolds embedded the data space, which allows for a meaningful way of data representing and interpreting. The following are important points and an example of challenges in medical imaging where manifold learning is used for identifying latent structure.

For example, a 3D medical imaging dataset might consist of both the coordinates (x,y,z) and the corresponding pixel intensities. The aim is to find a 2D-manifold in which the high-dimensional structure underlying the data lies.

Using Isomap:

- 1. Compute pairwise geodesic distances $dM(x_i, x_j)$ using a graph constructed from nearest neighbors.
- 2. Apply multidimensional scaling (MDS) to map the geodesic distances into a 2D Euclidean space.
- 3. Visualize the 2D representation to identify clusters or disease patterns.

This mathematical process enables researchers to uncover hidden structures, interpret data better, and make informed decisions in real-world healthcare scenarios.

4. Related Work

In this section, we present the state-of-the-art mathematical methods in AI for healthcare, compare different approaches that allow one to perform manifold learning, discuss the significance of differential geometry in the context of HD data analysis, and highlight limits of the known procedures to emphasize the importance of the present concept.

4.1 Comparative Analysis of Manifold Learning Techniques

There are several manifold learning algorithms, designed to deal with high-dimensional data as shown in Table 1:

Isometric Mapping (Isomap): Preserves global geometric structures by holding geodesic distances of data points. It provides an effective way to model the intrinsic manifold structure of the data, but it is computationally expensive especially when n is large [3].

Locally Linear Embedding (LLE): Emphasizes local neighboring relations, which is suitable for unwrapping complicated manifolds.

Table 1
Comparison of Manifold Learning Techniques

Technique	Preservation Focus	Computational Complexity	Scalability	Reference
Isomap	Global structures	High	Moderate	[3]
LLE	Local structures	Moderate	Low	[2]
t-SNE	Local structures	High	Low	[4]
UMAP	Local and global	Low	High	[4,5]

However, it may be highly sensitive to noise and demand careful choices of the neighborhood parameters [2].

t-Distributed Stochastic Neighbor Embedding (t-SNE): Focuses on local structure by transforming high-dimensional Euclidean distances to conditional probabilities. Good at visualizing local structure but may distort global structures and is computationally expensive [4].

Uniform Manifold Approximation and Projection (UMAP): Trade-off between preservation of local and global structure, faster computation and scalability. It has been demonstrated to generate embeddings similar to t-SNE and to be much faster to compute [4, 5].

4.2 Role of Differential Geometry in Understanding High-Dimensional Data

Differential geometry offers a method to study the geometric properties of high-dimensional data manifolds. It allows for better understanding of the intrinsic structure of data through geometric elements like curvature and geodesics [6]. This modeling technique is extremely valuable for healthcare where complex biological dynamics and/or diseases evolution pathways can be modeled. For example, differential geometry approaches have been used to design manifold-based neural networks that facilitate analysis on non-Euclidean manifolds in medical applications [7, 8].

4.3 Identified Gaps in Existing Methods and Relevance of the Proposed Conceptual Approach

The proposed conceptual approach Data on risk governance and management is collected through a survey that aims to assess different aspects of risk management, especially whether and how risks are addressed in policies and exercises at the local level and at the risk king level, and to assess whether and how specific measures to implement RSCs have been undertaken. Data on risk perception and acceptance of risk measures is collected through a survey that combines Revel'' scale with open (qualitative) questions to gather specific local perceptions and to assess the discussion on restoration policies. Also, the use of differential geometry in AI models of healthcare is not well-studied [9, 10, 11]. The conceptual methodology outlined here tries to tackle these challenges by utilizing manifold learning along with differential geometry to augment both disease detection and risk prediction and achieving a deeper understanding into complex medical dataset. Through the integration of manifold learning and differential geometry, the proposed framework

seeks to enhance the interpretability and awareness of AI models in healthcare, and to promote more efficient and personalized medical interventions [12].

5. Proposed Framework

This section introduces a conceptual framework that integrates manifold learning and differential geometry to analyze healthcare data. The proposed framework emphasizes the representation of high-dimensional medical data as manifolds, leveraging geometric properties for feature selection, disease detection, and subtype classification.

5.1 Healthcare Data as High-Dimensional Manifolds

Healthcare data, such as imaging datasets, electronic health records (EHRs), and genomic sequences, often exist in high-dimensional spaces ($\mathbb{R}^D$) due to the multitude of features involved. For instance, a single MRI scan can generate thousands of intensity values per voxel, while genomic data involves thousands of gene expressions per sample. These data are not uniformly distributed in ($\mathbb{R}^D$) but lie on lower-dimensional manifolds M that encapsulate the intrinsic structure of the data.

Mathematically, the manifold M is a subset of ($\mathbb{R}^D$) where each point $x \in M$ is associated with a local Euclidean neighborhood. Formally:

$$M \subset \mathbb{R}^D \text{ such that for all } x \in M, \exists U \subset M \text{ and } \phi: U \rightarrow \mathbb{R}^d$$

where $d \ll D$ and ϕ is a homeomorphism.

5.2 Approach for Clustering Disease Subtypes Based on Manifold Structures

Clustering disease subtypes involves identifying distinct regions on the manifold that correspond to specific patient groups. For example, t-SNE or UMAP can project patient data into a 2D or 3D space, where clusters represent different disease subtypes. Curvature analysis can further refine these clusters by identifying points of rapid change in data properties.

Algorithmically, clustering is achieved by:

1. Mapping high-dimensional data onto a low-dimensional manifold using manifold learning.
2. Applying clustering algorithms (e.g., k-means, DBSCAN) on the reduced representation.

- 3. Validating clusters using geometric measures such as intra-cluster geodesic distances.

This conceptual framework integrates manifold learning and differential geometry to provide interpretable and accurate insights for disease detection and subtype classification.

Key Steps as shown in figure 1:

- 1. Preprocessing:
Normalizes the data to ensure numerical stability and avoids biases in feature scaling.
Constructs a k-NN graph, which is crucial for manifold learning techniques like Isomap and UMAP.
- 2. Manifold Learning:
Maps high-dimensional data onto a low-dimensional space while preserving geometric structures.
Techniques such as Isomap, t-SNE, or UMAP optimize different properties, like geodesic distances or local neighbor relationships.
- 3. Geometric Analysis:
Uses differential geometry to compute curvature and geodesics. Curvature identifies points of rapid change (e.g., anomalies), and geodesics help model disease progression pathways.

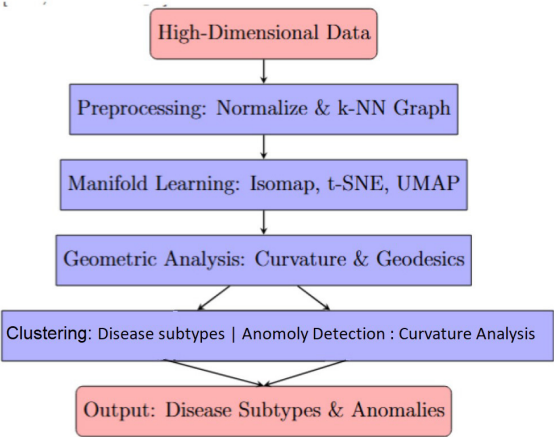


Figure 1
Workflow Architecture of proposed framework

4. Clustering:
Identifies disease subtypes by segmenting the low-dimensional manifold into clusters. This helps classify patients into distinct groups based on their health profiles.
5. Anomaly Detection:
Flags unusual patterns or rare diseases by analyzing deviations in curvature or proximity to cluster centroids.

Algorithm DiseaseDetectionManifoldLearning

Input: High-dimensional data: features

Output: Disease subtypes C , anomaly scores A

1. # Preprocessing
Normalize (X) , construct k-NN graph $(G(V, E))$, and compute pairwise distances.
2. # Manifold Learning
Apply a manifold learning method $f: \mathbb{R}^D$ to $\{\mathbb{R}\}^d$ ($d \ll D$):
- Isomap: Compute geodesic distances and use multidimensional scaling.
- t-SNE or UMAP: Optimize local/global neighborhood preservation.
3. Obtain low-dimensional embedding Z in $\{\mathbb{R}\}^{n \times d}$.
4. # Geometric Analysis
Compute geometric properties for (Z) :
- Curvature $(K(x_i))$: Detect variation or anomalies.
- Geodesic distances: Model relationships or progression pathways.
5. # Clustering and Anomaly Detection
6. Cluster (Z) using k-means or DBSCAN to identify disease subtypes (C) .
7. Compute anomaly scores (A) based on curvature (K) or cluster distances.
8. Return: Disease subtypes (C) , anomaly scores (A) .

This pseudo-code outlines a proposed framework for leveraging manifold learning and differential geometry to detect disease subtypes and identify anomalies in high-dimensional healthcare data. It begins with data preprocessing, including normalization and constructing a

k-nearest neighbors (k-NN) graph to define relationships between data points. Manifold learning techniques like Isomap, t-SNE, or UMAP are then applied to reduce the data's dimensionality while preserving intrinsic geometric properties. The reduced representation is analyzed using geometric measures such as curvature and geodesic distances, which help uncover latent structures and progression pathways. Finally, clustering algorithms (e.g., k-means or DBSCAN) are employed to classify data into disease subtypes, while anomalies are flagged based on curvature deviations or distances from cluster centroids, enabling robust disease modeling and risk prediction.

6. Conclusion

This integration of manifold learning and differential geometry will facilitate studies involving high-dimensional healthcare data. With the aid of manifold learning methods such as Isomap, t-SNE and UMAP, the framework can effectively lower the dimensions while retaining the intrinsic geometric structure, which makes it easier to find the hidden patterns. Differential geometry further enhances this analysis with constructions such as curvature and geodesics, useful for informative relationships and uncovering hitherto latent information in the data. The proposed model provides a reliable method of disease detection and subtype classification while keeping interpretability and precision for healthcare applications. The ability to capture local and global features of data allows to represent complex disease progression patterns and detect rare diseases with high confidence. The integration of AI and advanced mathematical concepts such as this not only enriches the analysis capability but will facilitate further studies related to personalized medicine, real time monitoring and scalable health care solutions.

References

- [1] J. Hu, R. Lepore, R. J. B. Dobson, A. M. Al-Chalabi, D. Bean, and A. Iacoangeli, "DGLinker: Flexible knowledge-graph prediction of disease–gene associations," *Nucleic Acids Res.*, vol. 49, no. W1, pp. W153–W161 (2021).
- [2] J. Vilela, M. Asif, A. R. Marques, J. X. Santos, C. Rasga, A. Vicente, H. Martiniano, "Biomedical knowledge graph embeddings for personalized medicine: Predicting disease–gene associations," *Expert Syst. Appl.*, vol. 40, e13181 (2023).

- [3] P. Cinaglia and M. Cannataro, "Identifying candidate gene–disease associations via graph neural networks," *Entropy*, vol. 25, no. 6, pp. 909 (2023).
- [4] M.-L. Zhang, B.-W. Zhao, P.-W. Hu, L. Wong, X.-R. Su, B.-Q. Wang, W. Liu, X.-Y. Zhang, and Z.-H. Zhang, "RLFDDA: A meta-path-based graph representation learning model for drug–disease association prediction," *BMC Bioinformatics*, vol. 23, no. 1, pp. 516 (2022).
- [5] K. Syama, J. Arul Jothi, and N. Khanna, "Automatic disease prediction from human gut metagenomic data using boosting Graph-SAGE," *BMC Bioinformatics*, vol. 24, pp. 126 (2023).
- [6] S. Zheng, Z. Zhu, Z. Liu, Z. Guo, Y. Liu, Y. Yang, and Y. Zhao, "Multi-modal graph learning for disease prediction," *IEEE Trans. Med. Imaging*, vol. 41, no. 9, pp. 2207–2216 (2022).
- [7] R. Guo, X. Tian, H. Lin, S. McKenna, H.-D. Li, F. Guo, and J. Liu, "Graph-based fusion of imaging, genetic and clinical data for degenerative disease diagnosis," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 21, no. 1, pp. 57–68 (2024).
- [8] Y. Yang, C. Ye, X. Guo, T. Wu, Y. Xiang, and T. Ma, "Mapping multi-modal brain connectome for brain disorder diagnosis via cross-modal mutual learning," *IEEE Trans. Med. Imaging*, vol. 43, no. 1, pp. 108–121 (2024).
- [9] R. Saha, D. K. Saha, Z. Fu, R. F. Silva, and V. D. Calhoun, "Multimodal fusion of functional and structural data to recognize longitudinal change patterns in the adolescent brain," in *Proc. 2023 IEEE EMBS Int. Conf. Biomed. Health Inform. (BHI)*, pp. 1–5 (2023).
- [10] C. Leelavathi, G. Thahniyath and K. R. Prasad, "Optimizing the deep features using the spectral concept for an efficient video classification," *Journal of Information and Optimization Sciences*, vol. 46, no. 2, pp. 563–562.
- [11] Y. Li, X. Zhang, J. Nie, G. Zhang, R. Fang, X. Xu, Z. Wu, D. Hu, L. Wang, H. Zhang, W. Lin, and G. Li, "Brain connectivity-based graph convolutional networks and its application to infant age prediction," *IEEE Trans. Med. Imaging*, vol. 41, no. 10, pp. 2764–2776 (2022).
- [12] S. Miri and M. Ilmakchi, "The scalar curvature of CR submanifold of maximal CR dimension with a conformal change," *Journal of Dynamical Systems and Geometric Theories*, vol. 21, no.2, pp. 201–214.

Received December, 2024