Quest Journals
Journal of Research in Applied Mathematics
Volume 10 ~ Issue 9 (2024) pp: 09-14
ISSN (Online): 2394-0743 ISSN (Print): 2394-0735
www.questjournals.org

Review Paper

Topological Approaches to Diabetes Prediction Using TDA

Revathi G [1]* and Gnanambal Ilango[2]
Research Scholar, Department of Mathematics, Government Arts College(A),
CBE-18
Associate Professor, Department of Mathematics, Government Arts College(A)
CBE-18

Abstract:
The concept of shape and significance in data is fundamental to Topological Data Analysis (TDA). This method has increased process efficiency across multiple sectors, including computational biology and healthcare. Utilizing the Kaggle Pima Indian Diabetes Dataset, persistent homology and persistent landscape were used in the present study to predict diabetes, demonstrating the significance of TDA in the medical field. To guide the investigation, we determined the persistent homology of a point cloud represented by a Vietoris-Rips complex. Additionally, we used the persistent landscape to identify six important features in the dataset. Consequently, we found that our prediction accuracy for diabetes is 75%.

Keywords: *Topological data analysis, Persistent diagram, Persistent landscape*

***R**eceived 05 Sep., 2024; **R**evised 15 Sep., 2024; **A**ccepted 18 Sep., 2024 © The author(s) 2024.*
P**ublished with open access at **www.questjournas.org

I. Introduction

Topological Data Analysis (TDA) is a rapidly evolving field that leverages concepts from algebraic topology to study the shape of data. The core idea behind TDA is that data possesses an inherent form and structure, which can be systematically analyzed and interpreted. One of the key tools in TDA is persistent homology, which captures topological features of data across multiple scales. This technique has found applications in various domains, including computational biology, healthcare, and more. The theoretical underpinnings of TDA are well-documented in foundational texts such as Computational Topology[5]. This book provides a comprehensive introduction to the field, covering essential concepts such as simplicial complexes, homology, and persistent homology.

One of the primary motivations for using TDA in healthcare is its ability to handle complex, high-dimensional data. In the context of medical data analysis, persistent homology offers a unique way to identify and quantify features that traditional statistical methods might overlook.

Recent advances in TDA have focused on making the techniques more accessible and practical for data scientists. Chazal and Michel (2021)[4] provide an accessible introduction to TDA, emphasizing both fundamental and practical aspects. Their work serves as a bridge between theoretical developments and practical applications, highlighting tools and methodologies that can be directly applied to real-world datasets.

A notable advancement in TDA is the concept of persistence landscapes, which provide a robust summary of persistence diagrams. Bubenik (2015) [2] introduced persistence landscapes as a tool to convert persistence diagrams into a functional space, enabling the use of statistical methods directly on these summaries. In embedded systems with limited resources, persistence landscapes have shown to be especially helpful for implementing dataset verification techniques. By transforming topological features into a format suitable for standard data analysis techniques, persistence landscapes facilitate the integration of TDA in various computational environments, including those with limited resources.

This paper aims to demonstrate the utility of TDA in the medical field, specifically in predicting diabetes using persistent homology and persistent landscape. By applying this technique to the Kaggle Pima Indian diabetes dataset, we aim to highlight the efficacy of persistent homology in identifying significant patterns and improving prediction accuracy.

II. Methodology

1. Dataset and Feature Selection

The Pima Indian Diabetes Dataset https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data [13] is a popular dataset used in the field of machine learning and data analysis, particularly for developing and testing algorithms for predicting diabetes. It is named after the Pima Indian population in the United States, which has a high prevalence of diabetes, making it an ideal candidate for studying this disease. The National Institute of Diabetes and Digestive and Kidney Diseases initially assembled the dataset, which the UCI Machine Learning Repository then made accessible to the general public. It was hosted on Kaggle, a well-known website for datasets and competitions in data science.

The collection has 768 samples, each of which corresponds to a female patient with Pima Indian ancestry. There are eight feature variables and a single target variable in every sample.

2. Persistent Homology Calculation

To compute the persistent homology, we represented the dataset as a point cloud and constructed a Vietoris-Rips complex. This complex captures the topological features of the data by connecting points that are within a certain distance of each other. Persistent homology tracks these features across different scales, providing a multi-scale view of the data's topological structure.

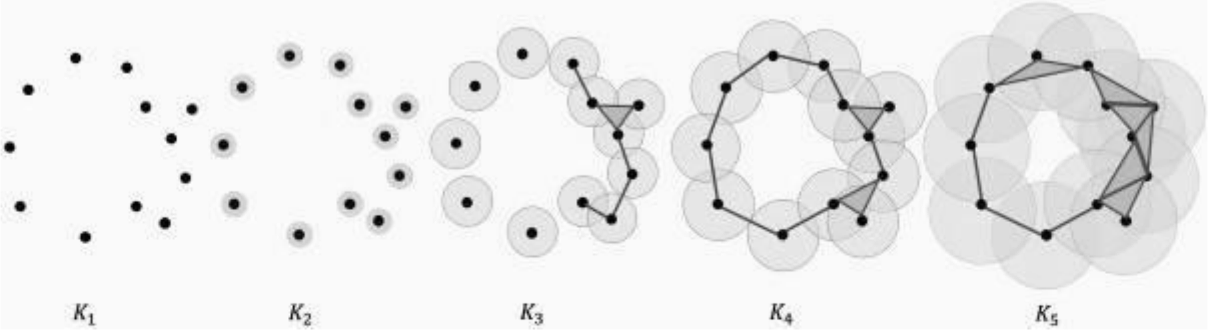


Figure 1: Filtration with the sequence of simplicial complex

The tracked features (birth, death) are recorded and visualize in many ways. The most ways are persistent diagrams and persistent barcodes.

2.1. Filtration

Let K "be a finite simplicial complex and $K_1 \subseteq K_2 \subseteq \cdots \subseteq K_r = K$ be a finite sequence of nested subcomplexes of K. K is called a **filtered simplicial complex** and the sequence $\{K_1, K_2, \dots\}$ is called a **filtration** of K" [1].

Numerous filtered simplicial complexes exist, including the Cech, Alpha, and Delaunay types. The Vietoris-Rips complex is our choice due of its computational efficiency.

Let $X = \{x_1, x_2, \dots . x_n\}$ be a collection of points in R^d. Given distance > 0 , R(X , ϵ) denotes the simplicial complex on n vertices $x_1, x_2, \dots . x_n$, where an edge between the vertices x_i and x_j with $i \neq j$ is included if and only if $d(x_i, x_j) \leq \epsilon$. This type of simplicial complex is called a **Vietoris-Rips complex** [4]**.**

Using Vietoris_Rips complex, the topological features are tracked and recorded in the diagram. This diagram is called Persistent Diagram.

2.2. Persistent Diagrams

The sets of topological pairings (birth, death) that emerge from filtering a simplicial complex are known as persistent diagrams. In this case, the pairings (b_i, d_j) denote the birth and the death, respectively. This collection is multi set, a set the elements can appear multiple times. Both the birth and death times of loops are shared by each pair, which is shown by their multiplicity. Stability is by far PDs' most important feature. The associated PD changes little in response to small modifications to the simplicial complex. In terms of applications, this feature is essential.

It guarantees repeatability and resistance against noise. Multi sets cannot be directly processed by a machine learning algorithm because they lack the key mathematical and statistical features needed in a machine learning setting. Therefore, in order to portray the persistent diagram as a vector and represent the diagram, a transformation is required. We refer to these changes as vectorization.

2.3. Vectorization methods

Persistent Landscape

The Persistent Landscape is one of the vectorization methods of representation of persistent diagram that gives the statistical properties in [2]. Formally, persistent Landscapes are piecewise constant functions $\lambda\colon \mathbb{N} \times \mathbb{R} \to \overline{\mathbb{R}}$. To define λ, we tent each persistent point $q = \left(\frac{b+d}{2}, \frac{d-b}{2}\right) \in D$ to the baseline $x = 0$ with the following function

$$\Lambda_q(t) = \begin{cases} t - b & \text{if } t \in \left[b, \frac{b+d}{2}\right], \\ d - t & \text{if } t \in \left[\frac{b+d}{2}, d\right], \\ 0 & \text{otherwise.} \end{cases}$$

The persistent Landscape of D is the collection of such functions $\lambda_D(k,t) = kmax_{q\epsilon D}\Lambda_q(t), \quad k\,\epsilon\mathbb{N}, t\epsilon[0,T]$,................................ (1)

where *kmax* is the kth largest element in the set and T is the real number such that $d \leq T$ for any death time d of a topological feature [2].

2.4. Machine learning Classifier

We are able to show the vectorization of Persistent Diagrams in machine learning classifiers at last. Machine learning classifiers are algorithms that use input feature-based classification to group data into predetermined classes or labels. Many applications, including sentiment analysis, picture recognition, spam detection, and medical diagnosis, depend on these classifiers. The Random Forest Classifier was employed in this study.

A strong and adaptable algorithm, Random Forest can handle a broad variety of data types and applications. It's a well-liked option in the machine learning field since it combines the advantages of several decision trees to produce reliable and accurate predictions.

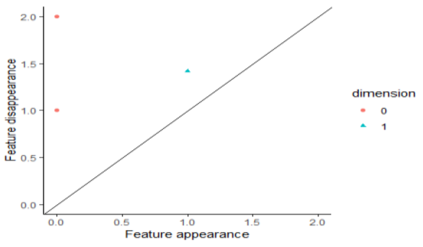
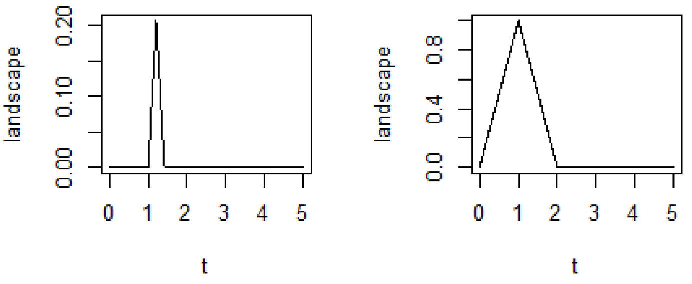
III. Data Classification

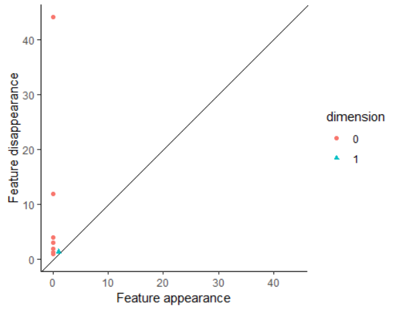
3.1. TDA method

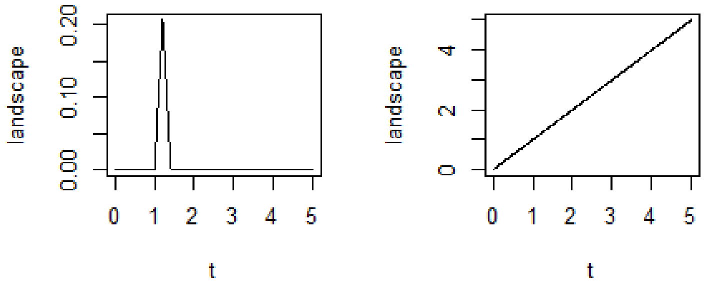
The Persistent homology is used for data projections into a Persistent diagram. It is built using R Studio. It utilizes the programming language R, free Integrated Development Environment (IDE) [6].

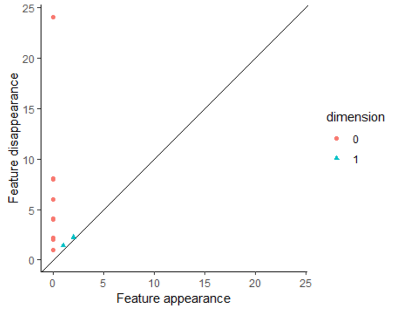
3.2. Data preprocessing

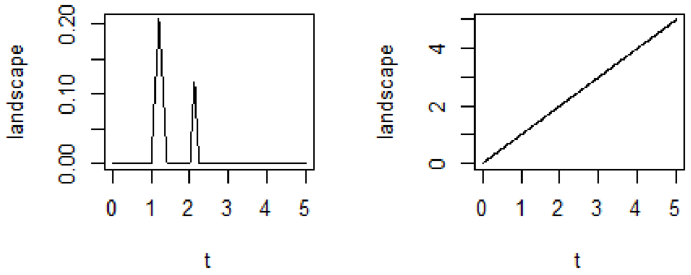
The Diabetes data collection contains no missing values. In the outcome Column “1” indicates the presence of Diabetes and “0” indicates no Diabetes. For each dataset, we calculated the Euclidean distance matrix, which represents the pairwise distances between data points. Using [6], the function (calculate_homology) computes the Vietoris-Rips complexes for each dataset. It represented by the distance matrix, helping to identify topological features like connected components and loops. These features are represented in the persistent diagram.

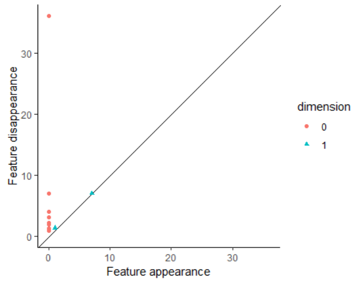
Persistent Diagram	**Persistent landscape (dim 1, dim 0)**
	

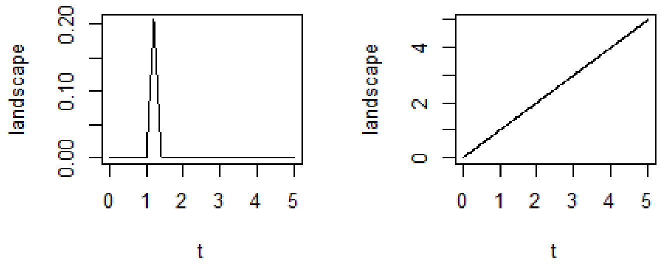

Feature disappearance
Feature appearance
dimension
0
1

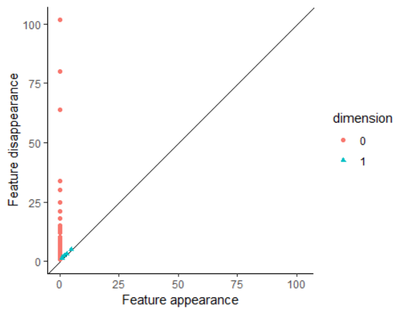

landscape
t
landscape
t

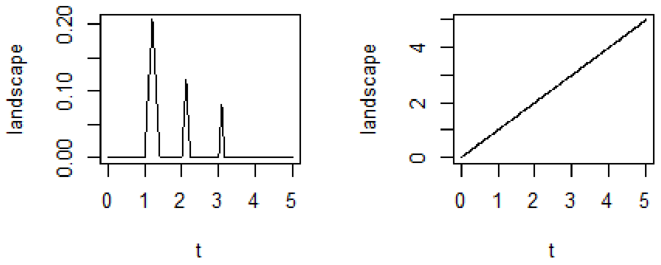

Feature disappearance
Feature appearance
dimension
0
1

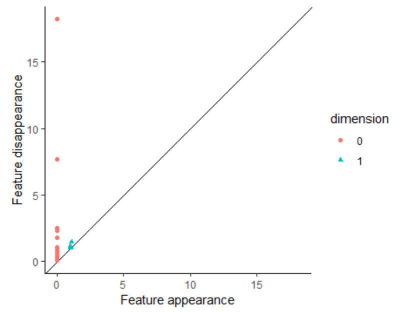
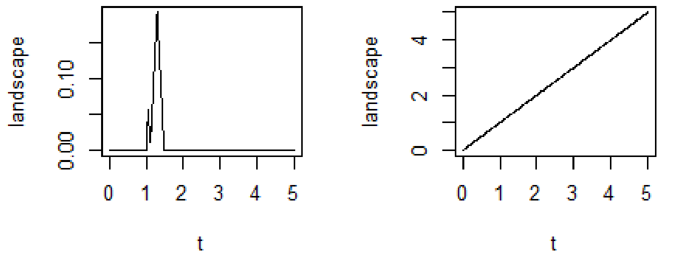
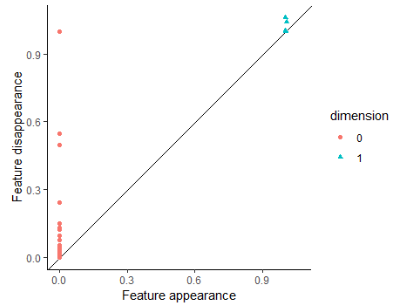
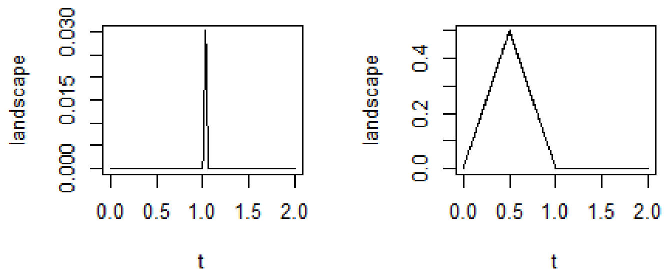
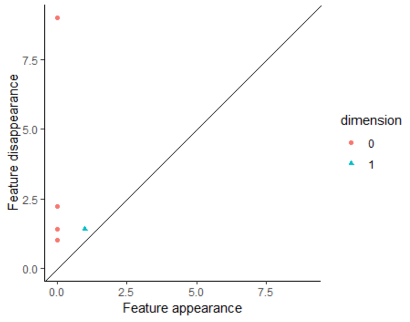
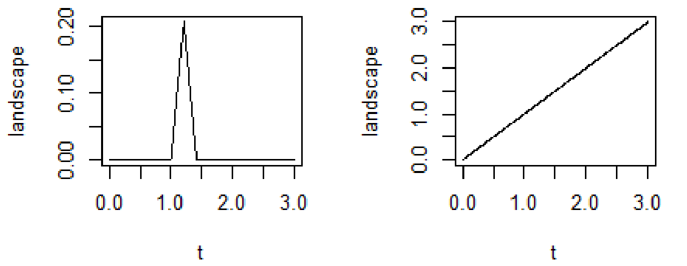

landscape
t
landscape
t


Feature disappearance
Feature appearance
dimension
0
1


landscape
t
landscape
t


Feature disappearance
Feature appearance
dimension
0
1


landscape
t
landscape
t

Also the function (landscape) [6] , the landscape function corresponding to a given persistence diagram is used .

Persistent Diagram, Persistence landscapes for dimension 0 (connected components) and dimension 1 (loops) for each dataset over a specified range of threshold values are given in the Table 1.

From (1), maximum Persistent landscape values are computed for each feature. Out of nine features six features are selected for the machine learning evaluation.

By the experiment the Blood Pressure, Insulin, Skin thickness, Age, Glucose, BMI are the key features for the diabetes

Data Evaluation

When we choose the nine features for the Random Forest Classifier, we can see that our suggested model performs noticeably better. During the assessment, the highest accuracy rate for Pima Indian diabetes dataset was 75% [3].

Analysis and Prediction

The primary objective was to leverage the topological features derived from persistent homology to predict diabetes. We used various machine learning models, concentrating on the accuracy of the predictions. The persistent homology features were used as inputs to these models, which were then trained and validated using the dataset.

IV. Results

Our analysis yielded a remarkable prediction accuracy of 75%. This high accuracy demonstrates the potential of persistent homology in extracting meaningful topological signatures from medical data, leading to highly accurate predictions.

V. Discussion

The results of this study underscore the effectiveness of Topological Data Analysis, and specifically persistent homology, in the medical field. By transforming the dataset into a topological space, we could uncover patterns that traditional methods might miss. The high prediction accuracy indicates that persistent homology can serve as a valuable tool in medical diagnostics and prognosis, particularly in the context of diabetes. Both results also take into account the algorithm that has the second-highest accuracy rate out of all the ones that were run.

VI. Conclusion

This research highlights the significance of Topological Data Analysis and persistent homology in enhancing the prediction of medical outcomes. The application of these techniques to the Pima Indian diabetes dataset resulted in a highly accurate predictive model. Future research could explore the integration of persistent homology with other advanced machine learning techniques to further improve prediction accuracy and applicability to other medical conditions.

Reference

[1]. Aguilar, Alejandro, and Katherine Ensor. "Topology data analysis using mean persistence landscapes in financial crashes." Journal of Mathematical Finance 10.4 (2020): 648-678.

[2]. Branco, Sérgio, et al. "Persistence Landscapes—Implementing a Dataset Verification Method in Resource-Scarce Embedded Systems." Computers 12.6 (2023): 110.

[3]. Bubenik, Peter. "Statistical topological data analysis using persistence landscapes." J. Mach. Learn. Res. 16.1 (2015): 77-102.

[4]. Bubenik, Peter, and Paweł Dłotko. "A persistence landscapes toolbox for topological statistics." Journal of Symbolic Computation 78 (2017): 91-114.

[5]. Chazal, Frédéric, and Bertrand Michel. "An introduction to topological data analysis: fundamental and practical aspects for data scientists." Frontiers in artificial intelligence 4 (2021): 667963.

[6]. Chang, Victor, et al. "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms." Neural Computing and Applications 35.22 (2023): 16157-16173.

[7]. Edelsbrunner, Herbert, and John L. Harer. Computational topology: an introduction. American Mathematical Society, 2022.

[8]. Fasy, Brittany Terese, et al. "Introduction to the R package TDA." arXiv preprint arXiv:1411.1830 (2014).

[9]. Nicolau, Monica, Arnold J. Levine, and Gunnar Carlsson. "Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival." Proceedings of the National Academy of Sciences 108.17 (2011): 7265-7270.

[10]. Somasundaram, Eashwar V., et al. "Benchmarking r packages for calculation of persistent homology." The R journal 13.1 (2021): 184.

[11]. Shultz, Christopher. "Applications of Topological Data Analysis in Economics." Available at SSRN 4378151 (2023).

[12]. Singh, Yashbir, et al. "Topological data analysis in medical imaging: current state of the art." Insights into Imaging 14.1 (2023): 58.

[13]. Data source: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database