International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(4s) 2026 pp. 216-226
https://doi.org/10.70917/ijcisim-2026-2505
Article

Persistent Homology Based Computational Data Analysis On Diabetes Mellitus

Sasikala D.[1], Abinaya S.[2]

[1]Department of Mathematics, PSGR Krishnammal College for Women, Coimbatore, Tamil Nadu, India.
Email: dsasikala@psgrkcw.ac.in
[2]Department of Mathematics, PSGR Krishnammal College for Women, Coimbatore, Tamil Nadu, India.
Email: abinaya170197@gmail.com

Abstract: The recent evolution in the dimensionality and volume of data makes it more difficult to examine and comprehend the findings. To encounter this, we use Topological Data Analysis (TDA), a paradigm for data analysis that identifies the geometrical structure of data and provides an interpretation based on the data's topological properties. In this paper, persistence diagrams of data from female patients with and without diabetes are computed and compared using bottleneck distance and Wasserstein distance to identify the less significant attributes that lead to the development of diabetes mellitus.
Keywords: NA

1. INTRODUCTION

Everything we interact with in this technological environment is centered around data. Data is used to train models and predict the results to enhance customer satisfaction. In medical industry, training data sets give fruitful results in early prediction of complications. The amount of data available is growing rapidly and is of high dimension. Analyzing the data has become more challenging due to two primary factors, the increasing magnitude and the complexity of the data which makes it difficult for the data scientists to deal with high dimensional noisy data. The existing clustering techniques and machine learning algorithms along with mathematical and statistical tools are becoming less effective. The fundamental problem is now to reduce noise and determine significant qualitative features of the data to draw necessary conclusions. Recent developments in mathematical research permit us to characterize a model for the data which proves to encounter the challenges faced.

Topological data analysis (TDA) is an implementation of topology in existing data analysis methods, to identify the geometric structure of data (Carlsson, 2009; Epstein et al., 2011) [1]. TDA links computer science, statistics, computational geometry, algebraic topology, data science and other related areas. Persistent homology is a concept evolved from homology in algebraic topology. This emerged through the work on Morse theory by Sergey Barannikov (1994) [2]. In this paper, the "birth-death" pairs were created from the canonically partitioned set containing critical values of the “smooth Morse function”. Filtered complexes were then segregated and their invariants, which are equivalent to persistence diagrams and barcodes, were also described along with an efficient algorithm for calculating them. One of the fundamental algorithms of TDA, Persistent homology [3], studies the shapes of data sets by creating a sequence involving simplicial complexes, and then computing the homology groups of each complex. Edelsbrunner H, et al., (2002) [4] proposed the algorithm for computation of persistent homology over , first of its kind. Later Zomorodian et al., (2005) [5] developed the computation of persistence homology over . Carlsson et al. (2009) [1] amended the preliminary definition and gave an analogous method of visualization called persistence barcodes, where persistence was interpreted in the language of commutative algebra. Cohen-Steiner D, et al., (2010) [6] used bottleneck distance as a measure to compare two persistence diagrams. In [6] the stability of persistent features under perturbation of the underlying data set is also proved. The first software package developed to compute persistence homology was JAVAPLEX (2014) [7]. Maria C, et al (2014) [8] developed GUDHI library for computation of persistence homology.

In this paper, we determine the less significant feature among the listed attributes of real time data set obtained from female diabetic and non-diabetic patients. We attempt to select the features that are similar for patients with and without diabetes by comparing the persistence diagrams of the attributes using suitable distance metrics.

In Section 2, we have presented the preliminary definitions necessary for the study. In Section 3, we have discussed the methodology adopted and in Section 4 we have presented inferences based on the results obtained and discussion on future work.

2. PRELIMINARIES

DEFINITION 2.1 [9]

A collection K, of subsets of a set K_0, which are non-empty, is called a simplicial complex if it is such that $\{w\} \in K$ for all $\tau \subset \sigma$, $w \in K_0$ & $\sigma \in K$ implies that $\tau \in K$. The elements belonging to K_0 are called vertices of K, and the elements belonging to K are called simplices. A p-simplex denoted by K_p has cardinality as p+1 whose dimension is p.

DEFINITION 2.2 [10]

Let $S = \{x_0, x_1, x_2 ..., x_k\}$ be a set and σ be a k-simplex defined on S. A simplex τ defined by T contained in S is a face of σ and has σ as a coface. The relationship is denoted with $\sigma \geq \tau$ and $\tau \leq \sigma$.

DEFINITION 2.3 [9]

The quotient vector space, $H_n(K)$, for n ranging over 0, 1, 2…. is the n^{th} homology of a simplicial complex K, given by

$$H_n(K) := \text{Kernel } (d_n) / \text{Image } (d_{n+1})$$

DEFINITION 2.4 [9]

The *m^{th} Betti number* of *K* given by $\beta m\,(K)$ is the dimension of the quotient vector space $H_n(K)$, where

$$\beta m\,(K) := \dim H_n(K) = \dim \text{kernel } (d_n) - \dim \text{image } (d_{n+1})$$

The n-boundaries are elements in the image of d_{n+1} and elements in the kernel of d_n are called n-cycles

DEFINITION 2.5 [11]

A sequence involving simplicial complexes, $\{K_t\}_{t\in I}$, (for an index set I), satisfying $K_{t1} \subseteq K_{t2}$, whenever $t_1 \leq t_2$, is called a filtration.

DEFINITION 2.6 [9]

Let a simplicial complex which is filtered, say, $K_1 \subset K_2 \subset \ldots\ldots\ldots \subset K_t = K$ be given. The pair $(\{H_n(K_i)\} 1 \leq i \leq t, \{f_{i,j}\} 1 \leq i \leq j \leq t)$, where for all i, j $\in \{1,2,\ldots.t\}$ with $i \leq j$, is the n^{th} persistent homology of K where the linear maps $f_{i,j} : H_n(K_i) \rightarrow H_n(K_j)$ are the maps induced by the inclusion maps $K_i \rightarrow K_j$.

DEFINITION 2.7 [9]

Let (X, d) be the space with metric d. For a non-empty subset S of X and ε, positive real number, the Vietoris-Rips Complex $VR\varepsilon$ (S) is defined as

$$VR_{\varepsilon}(S) = \{\sigma \subseteq S \mid d(x, y) \leq 2\varepsilon \ \forall\, x, y \in \sigma\}$$

It is also denoted as S_{ε}.

DEFINITION 2.8 [11]

A filtered simplical complex, say $K_1 \subset K_2 \subset \ldots\ldots \subset K_l = K$ be given and n be an integer. The n-th persistence diagram, $\wp_n(\{K_i\})^m$ of the filtration, is the multiset consisting of n-dimension holes present in the filtration. In other words, the birth and death of a n-dimensional hole present in a filtration is given by the pair (b, d) of integers, where b and d are denotes the birth and death of a hole respectively.

DEFINITION 2.9 [12]

For two elements D_1 and D_2 in the space of persistence diagrams, the bottleneck distance is given by

$$W_\infty(D_1, D_2) = \|x - \gamma(x)\|_\infty$$

DEFINITION 2.10 [12]

The n th Wasserstein distance between the persistence diagrams D_1 and D_2 in the space of persistence diagrams is calculated as

$$d_n(D_1, D_2) =$$

where the infimum is taken over all bijections between D_1 and D_2.

3. METHODOLOGY

3.1 COLLECTION OF DATA

The data set of interest consists medical record of female patients diagnosed with Diabetes Mellitus. Diabetes Mellitus is a medical condition which arises due to inadequate production of insulin by pancreas or the inability to respond to the insulin produced by our body. Insulin is a vital hormone, responsible for converting the glucose from food into energy. Abnormal insulin production can result in a sudden increase in blood glucose levels, leading to various health complications such as kidney failure, retinopathy, cardiovascular diseases, and more. Survey results shows that approximately 10.5% of world adult population is affected by diabetes and the number is growing continuously at an alarming rate. This makes it necessary to analyze and understand the consequence of the disease early so as to prevent further health issues and ensure healthy living. The data set used in the analysis was Pima Indian Diabetes Dataset [13] obtained from OpenML, which is an online machine learning platform that has access to data sets collected from several disciplines. The Public Domain License is accessible for this adequately anonymized dataset. This dataset was initially acquired from the National Institute of Diabetes and Digestive and Kidney Diseases, USA. It does not contain any identifiable features of the patients. There were totally around 768 female patients, all of age above 21, whose response was recorded over 9 attributes which includes age, blood pressure, insulin resistance, BMI, Diabetes Pedigree Function, etc.

3.2 DESCRIPTION OF THE ATTRIBUTES

1) 'preg'- This characteristic provides the respondent's total number of pregnancies.

2) 'plas'- This characteristic provides the concentration of glucose concentration in blood after 120 minutes in an oral glucose tolerance test.

3) 'pres' - This attribute characterizes the Diastolic blood pressure (mm Hg), which captures the arterial pressure, when the heart rests between the beats.

4) 'skin'- The thickness of the triceps skin fold (mm) is indicated by this feature. Research indicates that elevated blood glucose levels, can cause thick, rough, and puffy skin particularly in the hands, fingers, and toes, over a prolonged period of observation.

5) 'insu'- This characteristic indicates the level of glucose concentration in blood (mu U/ml) after 2 hours of food consumption. This test aids in the diagnosis of hypoglycaemia and associated disorders.

6) 'mass'- This attribute provides information about the body mass index, commonly known as BMI of the respondent. It is a measure of body fat, based on the height and weight of the individual.

7) 'pedi'- This attribute characterizes the Diabetes Pedigree Function (DPF) which determines the probability of diabetes based on the age and family history of diabetes in the respondent. It is proved to be one of the causes for diabetes mellitus.

8) **'age'**- This attribute gives the respondents age in years. Age-related increases in insulin resistance and deterioration of pancreatic islet function put older persons at increased risk of developing type 2 diabetes.

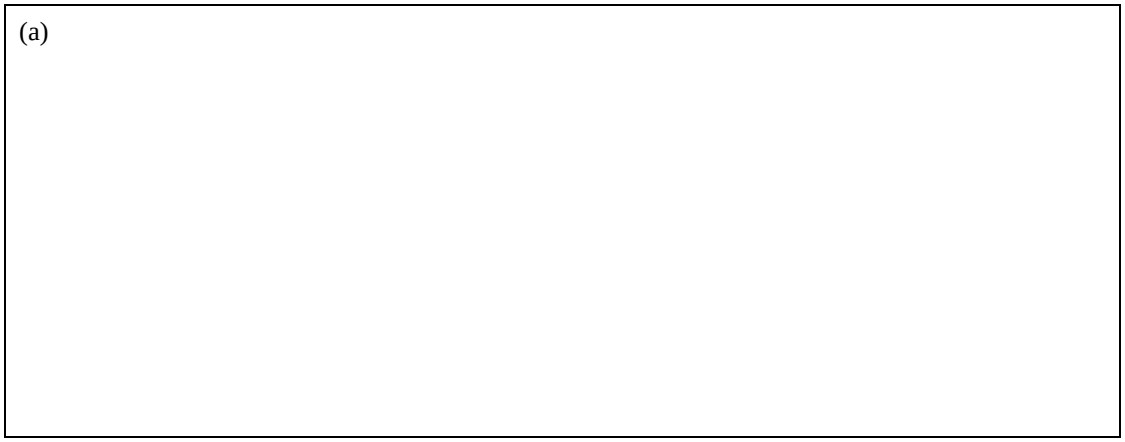
9) **'class'** - Tested positive/ negative of Diabetes Mellitus.

3.2 FILTRATION OF DATA

The raw data is very difficult to analyze and it does not provide valuable qualitative insights as such. The collected raw data has to be filtered into a sequence of simplicial complexes for further analysis. We have constructed Vietoris-Rips Complex to obtain a nested sequence of filtered simplicial complexes using GUDHI library. Among the several available complexes in TDA method, Vietoris-Rips Complex is considered to be more effective as it involves computation of complexes which are 0-dimensional and 1-dimensional only. This approach enables us to analyze how the shape of the data evolves across different resolutions, providing us with more meaningful information. Circles of chosen parameter as radius were drawn around each point. Two centers of circles were joined by an edge iff the Euclidean distance between them is not greater than the parameter value . We construct a triangle, if and only if all the faces of triangle lie in Similarly, we construct tetrahedron and other higher dimensional polytypes by varying the parameter value. The collection of all the points, edges, triangles, tetrahedron, etc., is called simplicial complex. By varying the parameter value , we can control the size of the balls and obtain a nested sequence involving filtered simplicial complexes. The optimal value of the scaling parameter chosen and the corresponding persistent diagram followed by persistent barcode was obtained. Fig 3.1 shows the simplicial complex obtained for diabetic and healthy patients without diabetes over the attributes.

3.3 PERSISTENT BARCODES AND DIAGRAMS

A persistent barcode is a finite multiset of points that are bounded below, marked along the real line. These are horizontal bars, which acts as a representative of the topological features of the data. The lifetime of the topological feature is obtained from the length of each bar represented from the filtered nested sequence of simplicial complex. The longer intervals relate to the strong topological features of the data while the shorter ones correspond to the noise in the data. Hence by inspecting the persistent barcodes we get clear information about the feature which persists throughout and the characteristics that will eventually disappear. We ignore the later and concentrate in identifying the one that persists throughout for further analysis. We have presented the persistence barcodes and persistence diagrams obtained by comparing the attributes as used earlier to compute simplicial complex. We have used GUDHI python module to compute the persistence barcodes and persistence diagrams. Persistent diagrams contain similar topological information of the data as persistent barcodes. The birth of a feature is indicated on the axis X, while its death is indicated on the axis Y. In a persistent diagram, the birth and death of a feature are denoted by the coordinates (b, d). A diagonal is drawn irrespective of the points obtained. As a feature cannot die before it is born, none of the points lie below the diagonal. The points in proximity to this diagonal are indicative of noise and are regarded as less significant. The points situated distant from the diagonal represent noteworthy features that require examination to reach precise conclusions. The dot in red colour represents the presence of connected component and the blue colored one indicates the 1-dimensional holes. For example, in fig 3.2 (a) there are two red dots far away from the diagonal indicating the presence of two connected components. On the other hand, no appreciable blue dot is located far from the diagonal, indicating the absence of a one-dimensional hole.



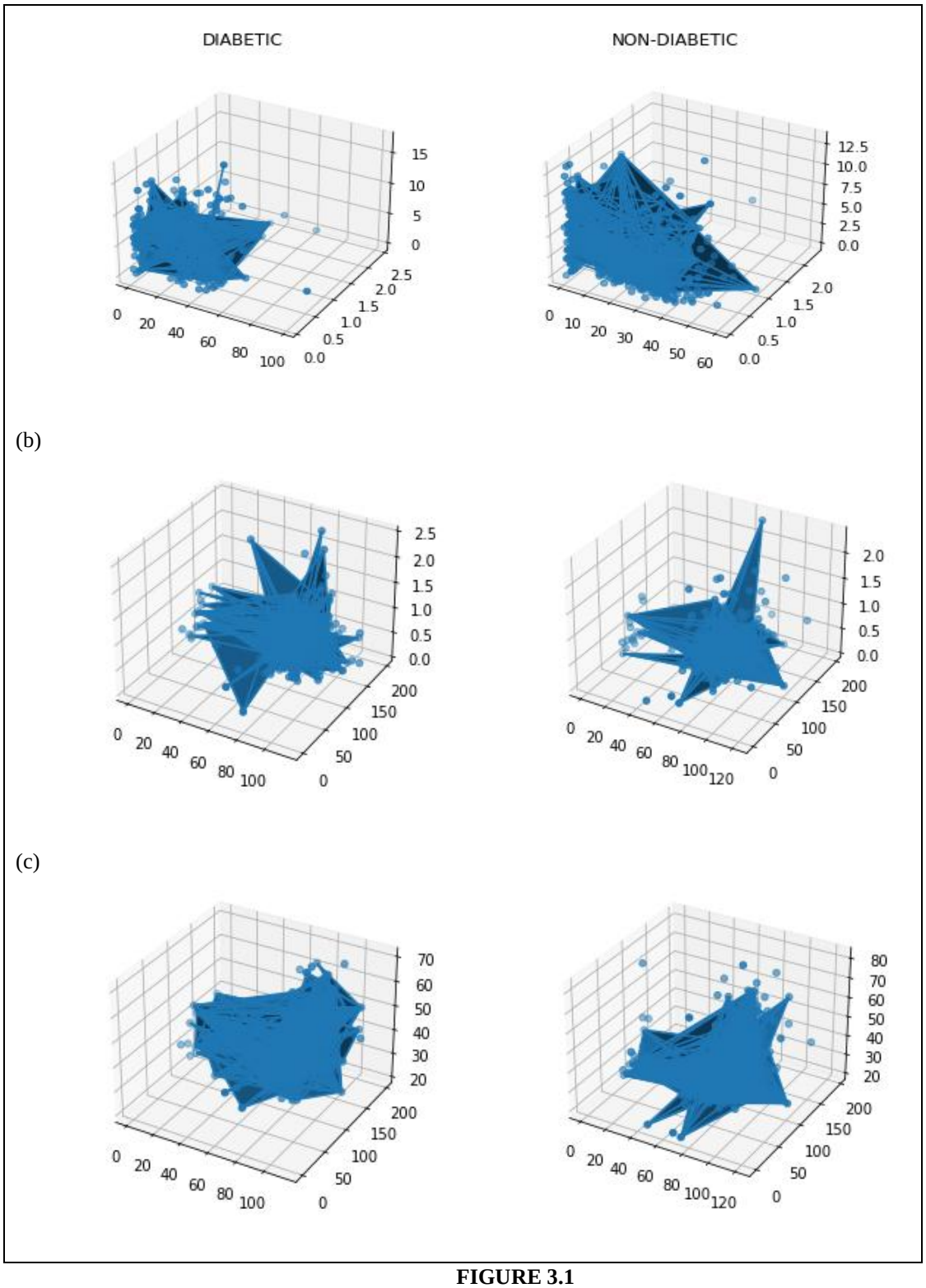
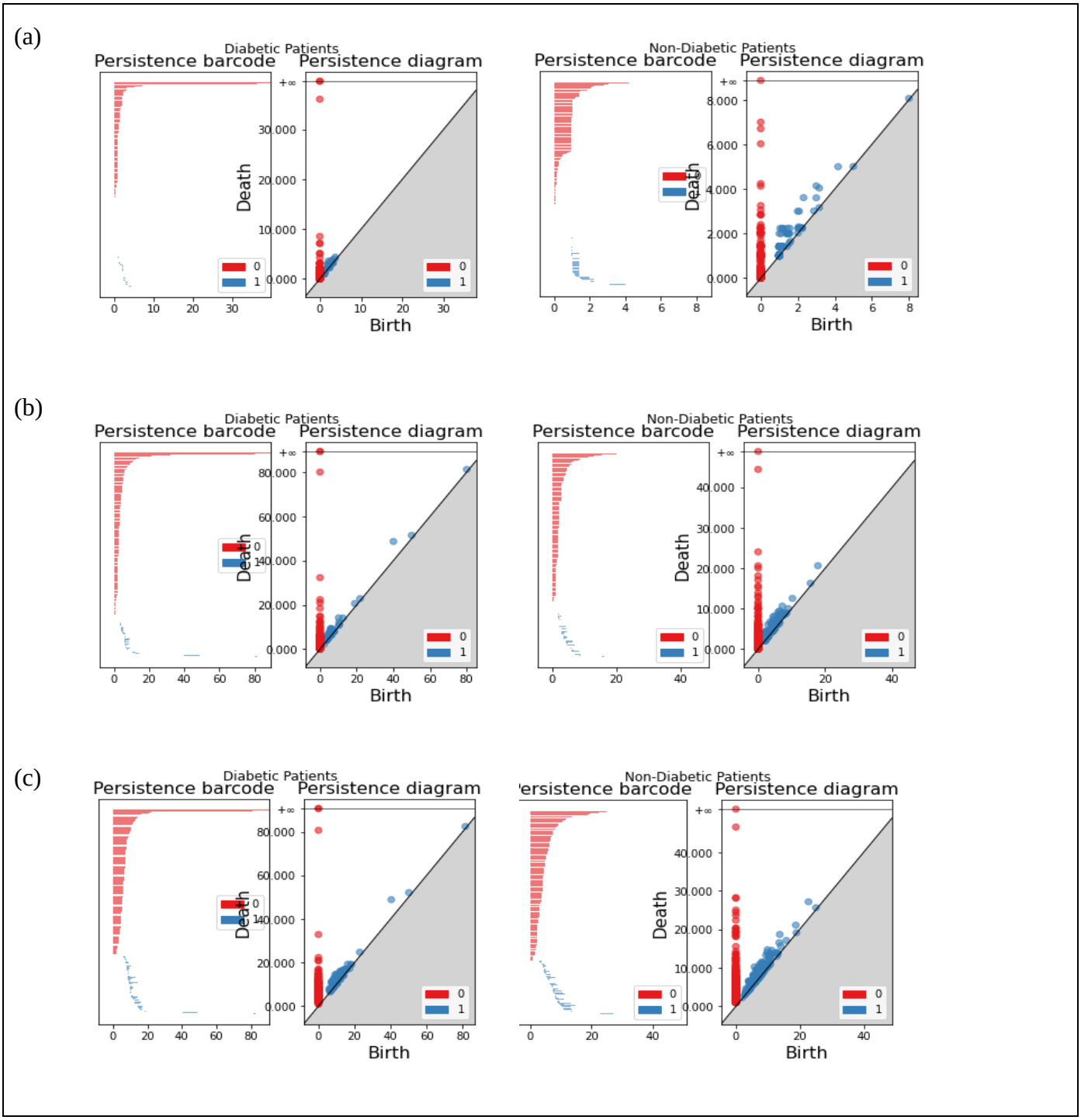


FIGURE 3.1

FIGURE 3.1 (a) Simplicial complex for Diabetic V/s healthy patients without diabetes for attributes “skin”, “pedi” and “preg”.

FIGURE 3.1 (b) Simplicial complex for Diabetic V/s healthy patients without diabetes for attributes “pres”, “plas” and “pedi”.

FIGURE 3.1 (c) Simplicial complex for Diabetic V/s healthy patients without diabetes for attributes “pres”, “plas” and “age”.



(d)

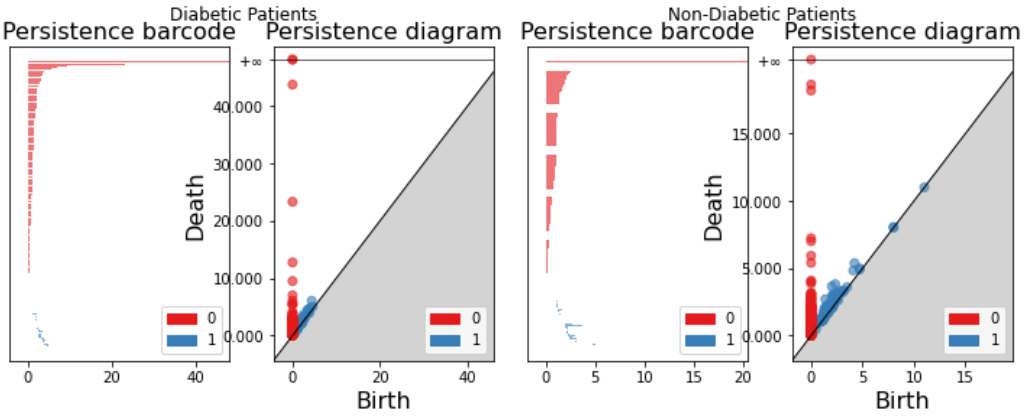


FIGURE 3.2

FIGURE 3.2 (a) Persistence Barcode V/s Persistence Diagram for Diabetic and Non-Diabetic patients for attributes “skin”, “pedi” and “preg”.

FIGURE 3.2 (b) Persistence Barcode V/s Persistence Diagram for Diabetic and Non-Diabetic patients for attributes “pres”, “plas” and “pedi”.

FIGURE 3.2 (c) Persistence Barcode V/s Persistence Diagram for Diabetic and Non-Diabetic patients for attributes “pres”, “plas” and “age”.

FIGURE 3.2 (d) Persistence Barcode V/s Persistence Diagram for Diabetic and Non-Diabetic patients for attributes “skin”, “pedi” and “mass”.

3.4 COMPARISON OF PERSISTENCE DIAGRAMS

The Wasserstein and bottleneck distances are the standard measures used to compare two persistence diagrams. Large values of these distances indicate that the diagrams are not likely similar and smaller values indicate that the persistence diagrams are similar. We have calculated bottleneck distance and Wasserstein distance for 1-dimensional persistence diagrams of diabetic and non-diabetic patients separately and some of them are listed in Table 3.1-3.4. We search for those attributes where the distances of persistence diagrams between patients with and without diabetes is minimal. Lesser value of distance implies greater degree of similarity in the persistence diagrams, which in turn makes these attributes less significant. In this regard, several combinations of attributes were chosen and their bottleneck distances and Wasserstein distances were computed and some of the key findings are highlighted in the table below.

TABLE 3.1 Values of Bottleneck Distance and Wasserstein Distance obtained for Persistence Diagrams of Diabetic and Non-Diabetic Patients by keeping “skin” fixed.

S.No	**Attribute**	**Bottleneck distance**	**1-Wassertein distance**	**2-Wasserstein distance**
	“skin”, *“preg”*	0.6180339887498949	20.39634394607585	2.462218771746149
	“skin”, *“plas”*	4.468667890309334	38.29830778698776	6.614790412670206
	“skin”, *“pres”*	1.8523499553598128	19.984101381705536	3.3637715959923637
	“skin”, *“insu”*	3.396078054371138	31.321580837890306	6.2887233475706505

	"skin", *"mass"*	0.7177427992306074	12.42936362739003	1.8770719484407457
	"skin", *"pedi"*	**0.0335277791041908**	**0.2707553462270368**	**0.0616479571169115 8**
	"skin", *"age"*	4.0	29.323526570032186	5.279952285194258

TABLE 3.2 Values of Bottleneck Distance and Wasserstein Distance obtained for Persistence Diagrams of Diabetic and Non-Diabetic Patients by keeping "skin" and "pedi" fixed.

S.No	Attribute	Bottleneck distance	1-Wassertein distance	2-Wasserstein distance
	"skin", "pedi", *"preg"*	**0.5361266967385567**	19.247960026645355	2.266991535021704
	"skin", "pedi", *"plas"*	4.465822810639743	38.53932346073205	6.601238847116559
	"skin", "pedi", *"pres"*	1.8509675733659705	21.02001576000781	3.3684545199049922
	"skin", "pedi", *"insu"*	3.375536762053539	31.401065358566164	6.24396014120105
	"skin", "pedi", *"mass"*	0.684225919030311	**12.538846144888245**	**1.7798212126815405**
	"skin", "pedi", *"age"*	3.987591157152826	29.12331697510563	5.243838165122927

TABLE 3.3 Values of Bottleneck Distance and Wasserstein Distance obtained for Persistence Diagrams of Diabetic and Non-Diabetic Patients by keeping "skin", "pedi" and "preg" fixed.

S.No	Attribute	Bottleneck distance	1-Wassertein distance	2-Wasserstein distance
	"skin", "pedi", "preg", *"plas"*	4.513720925768798	57.46475205074884	7.3145934504774 45
	"skin", "pedi", "preg", *"pres"*	1.5241792728300645	45.92442709996307	4.7304269324285615
	"skin", "pedi",	3.4440089711760535	41.601631958128685	6.593410999952291

	"preg", *"insu"*			
	"skin", "pedi", "preg", *"mass"*	**1.1845467702918997**	**28.972213561059146**	**3.1611138348004992**
	"skin", "pedi", "preg", *"age"*	2.8032983510559326	34.84636999442901	4.525738401500564

TABLE 3.4 Values of Bottleneck Distance and Wasserstein Distance obtained for Persistence Diagrams of Diabetic and Non-Diabetic Patients by keeping "skin", "pedi" and "preg" fixed.

S.No	Attribute	Bottleneck distance	1-Wassertein distance	2-Wasserstein distance
	"skin", "pedi", "mass", *"preg"*	**1.1845467702918988**	**28.972213561059146**	**3.1611138348004992**
	"skin", "pedi", "mass", *"plas"*	3.859751322491009	74.97946051527275	8.229374778942773
	"skin", "pedi", "mass", *"pres"*	1.6315704613941744	48.652518657629265	5.124655709337211
	"skin", "pedi", "mass", *"insu"*	3.550262829104174	49.94010402199482	7.917243800878374
	"skin", "pedi", "mass", *"age"*	3.122250765659916	48.392135486435244	6.546736589408641

4. RESULTS AND DISCUSSION

It was found that the distance was comparatively less for those diagrams involving the attribute "skin" which relates to skin thickness. We then proceed by keeping this attribute fixed and varying other attributes. Table 3.1 contains a tabulation of the values. We found that the attribute "skin" and "pedi" was least. We then fixed "skin" and "pedi" by varying others and the values are shown in table 3.2. As we reached forward, there came a point when we had two distinct parameters "preg" and "mass" with the lowest values. So, we again computed the distance by keeping "preg" fixed and varying others. The values are tabulated in table 3.3. It is seen that the feature "mass" has the least value. Upon keeping "mass" fixed, we had "preg" as the least value as shown in table 3.4. On proceeding further, the

Wasserstein distance and bottleneck grow, indicating a higher degree of dissimilarity. We stop at this juncture. After testing different combinations and fixing the attribute associated with the smallest distance, it can be inferred that skin thickness and the diabetes pedigree function were among the less impactful traits. The least value of the distance in each instance is highlighted.

5. CONCLUSION

An analysis on Pima Indian diabetes dataset is done by employing the concept of persistent homology. The dataset comprises medical information gathered from female patients with and without diabetes. Simplicial complex, persistence barcodes, persistence diagrams and their corresponding bottleneck distance and Wasserstein distance were computed for the attributes listed. Upon analysis, we were able to conclude that skin thickness and Diabetes Pedigree Function are among the less significant features that can be considered to determine diabetes mellitus. This approach offers a versatile tool for detecting and subsequently removing less consequential attributes within any multi-dimensional dataset. This streamlines the analysis process significantly, facilitating clearer insights and more precise conclusions. In future, we would like to develop a new model to diagnose the early development of the condition based on the listed attributes.

References

1. G. Carlsson (2009). Topology and Data. Bulletin of The American Mathematical Society - BULL AMER MATH SOC. 46. 255-308. 10.1090/S0273-0979-09-01249-X.
2. S. Barannikov. "The framed Morse complex and its invariants." Advances in Soviet Mathematics 21 (1994): 93-116.
3. H. Edelsbrunner, D. Morozov (2012) Persistent homology: theory and practice. In:Proceedings of the European congress of mathematics, pp 31-50.
4. H. Edelsbrunner, Letscher, and A. Zomorodian. "Topological persistence and simplification." Discrete & Computational Geometry 28 (2002): 511-533.
5. A. Zomorodian, G. Carlsson (2005), Computing persistent homology. Discrete Comput Geom 33:249-274.
6. Cohen-Steiner, David, et al. "Lipschitz functions have L p-stable persistence." Foundations of computational mathematics 10.2 (2010): 127-139.
7. A. Tausz, M. Vejdemo-Johansson, H. Adams (2014) JavaPlex: a research software package for persistent (co)homology. In: Hong H, Yap C (eds) Mathematical software - ICMS 2014. Lecture notes in computer science, vol 8592, pp 129-136. Software available at http://appliedtopology.github.io/javaplex/.
8. C. Maria, J. D Boissonnat, M.Glisse, & M. Yvinec, (2014). The gudhi library: Simplicial complexes and persistent homology. In Mathematical Software–ICMS 2014: 4th International Congress, Seoul, South Korea, August 5-9, 2014. Proceedings 4 (pp. 167-174). Springer Berlin Heidelberg.
9. N. Otter, M.A. Porter, U. Tillmann, et al. A roadmap for the computation of persistent homology. EPJ Data Sci. 6, 17 (2017).
10. P. Sekuloski, & V. Dimitrievska Ristovska, (2019). Application of persistent homology on bio-medical data–a case study. Mathematical Modeling, 3(4), 109-112.
11. H. Edelsbrunner and John L. Harer. Computational topology: an introduction. American Mathematical Society, 2022.
12. Kerber, Michael, Dmitriy Morozov, and Arnur Nigmetov. "Geometry helps to compare persistence diagrams." (2017): 1-20.
13. Markelle Kelly, Rachel Longjohn, Kolby Nottingham, The UCI Machine Learning Repository, https://archive.ics.uci.edu.
14. A. Hatcher (2002). Algebraic topology. Cambridge University Press.