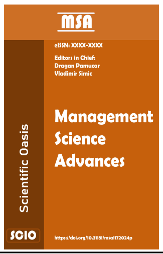
SCIENTIFIC OASIS

Management Science Advances

Journal homepage: www.msa-journal.org
ISSN: 3042-2205



msa

Management Science Advances

Scientific Oasis

SCIO

Optimizing Diabetes Data Insights Through Kmapper-Based Topological Networks: A Decision Analytics Approach for Predictive and Prescriptive Modeling

Muhammad Abid[1,*], Muhammad Saqlain[2]

[1] Department of Mathematics, North Carolina State University, Raleigh, 27606, NC, United States
[2] School of Mathematics, Northwest University (NWU), Xi'an 710069, Shaanxi, China.

ARTICLE INFO

Article history:
Received 14 October 2024
Received in revised form 15 November 2024
Accepted 17 November 2024
Available online 17 November 2024

ABSTRACT

A highly effective technique for identifying and showing structures in high-dimensional datasets is topological data analysis. The Kmapper software creates overlaying clustering graphs and topological networks to facilitate the investigation of such information. The objective of the work was to visualize a dataset of diabetes patients that included information on blood pressure, glucose levels, and pregnancies using the Kmapper software. Afterward, it applied topological data analysis to see if any underlying structures or patterns could be established. The preprocessed dataset of diabetic patients was acquired via Kaggle. Kmapper was run with a difference of parameter settings, which includes 0.4 overlap, 15 hypercubes, and varying numbers of PCA components (1, 2, and 3). We investigated the generated graph visualizations. Although two PCA components were used, the topological graphs disclosed likely intriguing highlights such as three peaks. To understand these illustrated structures in the background of the diabetes data, additional investigation is mandatory. Also with every aspect considered, Kmapper worked well for using topological representations to contribute intuition into the high-dimensional dataset.

Keywords:
Topological Data Analysis; Mapper Algorithm; Diabetes Patient Data; Machine Learning; High-Dimensional Data Visualization.

1. Introduction

As the high-dimensional and complex datasets can contain intrinsic structures and also the patterns that can be found using topological data analysis (TDA), a really strong technique [1]. TDA con-

[*] *Corresponding author.*
E-mail address: abidaminnaeem@gmail.com

https://doi.org/10.31181/msa1120241

tributes a advance method of investigating data by utilizing ideas from algebraic topology, which enables it possible to extract paramount insights that can be missed by more conventional techniques. This detailed overview investigates TDA, also including its theoretical foundations, real-world applications, and the most latest developments that have fueled its popularity in a diversification of fields [2].

The discipline of the mathematics of algebraic topology, which explores the characteristics of geometric objects that remain invariant despite continuous deformations, is the foundation of the field of TDA [3]. In a consequence of the distinct viewpoint, TDA is allowed to examine data as a geometric structure, revealing innate patterns and also shapes that would be challenging to identify using traditional statistical methods [4]. The field of TDA comes up with a strong framework for comprehending complex systems, ranging from social dynamics to biological networks, by summarizing the innate topology of data [5]. Persistent homology is an essential idea in TDA that offers a systematic process toward identifying and measuring topological properties across several scales, also including loops, voids, and connected components [6]. While working with noisy or incomplete data, this multiscale analysis is a more effective way since it may identify stable and enduring features while removing noise and sporadic patterns [7].

Implementation of TDA is established in many other fields, which including computer vision, materials science, healthcare, and network analysis [8]. TDA has been used in biomedicine to explore the architecture of biological networks, including brain connectomes [9] and also protein-protein interaction networks [10]. In addition, TDA has illustrated potential in comprehending the dynamics of disease progression [11] and discovering intrinsic patterns within cancer genomics data [12]. TDA has been used in research of the materials to describe and explain the intricate microstructures of many materials, allowing for the discovery of phase transitions, flaws, and hidden patterns that affect material characteristics [13]. Over and above, the TDA has been applied to computer vision problems like image segmentation [14], object identification [15], and shape analysis [16].

Additionally to its many other uses, TDA has seen substantial theoretical and algorithmic breakthroughs in the past [17]. The development of productive computational methods for persistent homology is one noteworthy area of progress that has made it possible to analyze large-scale datasets [18]. As topological machine learning (TML) has emerged as a result of research on the integration of TDA with machine learning methods [19]. Incorporating the best aspects of machine learning with TDA, TML allows topological characteristics to be extracted from data and used for a variety of learning tasks, including dimensionality reduction, clustering, and classification [20]. TML models have the potential to increase performance and illustratively by capturing intrinsic patterns and structures that typical machine learning techniques can miss by integrating topological information [21].

The creation of interactive and interpretable visualization tools is an intriguing emerging field of TDA research [22]. Researchers can give rise to user-friendly visualizations that make tough dataset exploration and comprehension easier by utilizing topological representations like persistence diagrams and the Mapper [23]. In fields of this kind as materials research, where they facilitate the identification of microstructural characteristics and their correlations with material properties, these visualizations have shown to be extremely useful [24]. The field of TDA has many benefits, but it also has certain disadvantages and difficulties. The computational difficulty of persistent homology computations is a major obstacle, particularly for high-dimensional datasets [25]. To solve this problem and enable scalable TDA applications, researchers are actively investigating new algorithms and approximation methods [26].

The interpretation and conversion of topological information into useful insights presents another difficulty [27]. Although TDA can reveal intrinsic structures and patterns, interpreting these characteristics frequently calls for domain-specific knowledge and experience [28]. To close the gap between topological representations and useful insights, ongoing research endeavors are focused on creating interpretability frameworks and domain-specific applications [29]. Researchers are perpetually investigating TDA's applicability in several developing sectors as it continues to gain popularity [30]. Analysis of dynamic and time-varying data is one area of specific interest, as the field of TDA can shed light on how topological properties change over time [31]. This has outstanding ramifications for domains where temporal patterns and transitions must be captured, such as financial data analysis, traffic monitoring, and video analysis [32–34].

Incorporating TDA with other data analysis methods, such as reinforcement learning and deep learning, is another exciting avenue [35]. Researchers hope to create hybrid models that can take dominance of TDA's topological comprehension and the expressive capacity of deep neural networks by merging the best features of both of these methods [36, 37]. These hybrid models could improve performance on tasks such as natural language processing, control systems, and image recognition [38]. The field of TDA has been comparatively difficult to adapt to industry and real-world applications, despite its many benefits. The interpretability and accessibility of TDA approaches for non-experts is one of the primary issues [39, 40]. To close the gap between theoretical TDA and real-world implementation, continuing efforts are being made to produce intuitive visualizations, user-friendly software tools, and domain-particular applications [41–43].

The creation of theoretical underpinnings and mathematical foundations that can harmonize and generalize the diverse TDA techniques is once more an area of significance. This requires developing new algebraic and categorical frameworks for encapsulating topological properties [44], as well as investigating the replacement concepts of persistence, such as zigzag persistence [45, 46]. Knowledge sharing and multidisciplinary cooperation among scholars in other fields will be crucial as TDA develops further [47]. Researchers can advantage of domain-specific knowledge and insights by promoting cross-disciplinary partnerships, which will eventually result in more reliable and outstanding TDA applications across a range of domains [48–50].

To encapsulate, although topological data analysis (TDA) is becoming more and more popular, there are still a lot of unanswered questions about how to use TDA methods to handle complicated, high-dimensional datasets from certain areas. Since many TDA approaches already in use are generic, they might not accurately represent the subtleties and distinctive qualities of data from certain disciplines [51–55]. The objective of this current study is to close this gap by creating a customized TDA method meant for the analysis of patient data related to diabetes. The Mapper algorithm is a vigorous visualization tool in TDA, and its creative application to reveal topological structures and patterns within the multidimensional data of diabetic patients is the main contribution. Between careful lens selection, domain-specific clustering techniques, and parameter choices, our approach recovers remarkable insights that would be demanding to gain through conventional analytical methods [56–58]. Recent studies explore advanced mathematical models in fluid dynamics, chaotic systems, and immune responses. [59]. address well-posedness in micropolar fluid equations, [60]. analyze fractal dynamics in UAVs, and [61]. model tumor-immune interactions using fractional derivatives.

This work is new in two different ways: firstly, it shows how the field of TDA may be used to reveal hidden patterns in intricate medical data, opening the door to better risk assessment, patient classification, and tailored treatment plans. Secondly, it suggests a foundation for customizing TDA methods to inscription issues, especially to a given domain, developing cross-disciplinary collaboration and also knowledge sharing in the middle of the domain experts and data analysts. So conclusively, our work in this paper highlights the significance of including topological viewpoints into pipelines for data analysis, providing a high-yielding lens through which to investigate the complex geometries that underpin complex systems.

2. Mathematical Formulations

The Mapper algorithm and the topological data analysis (TDA) are established on several mathematical recommendations and also include the methods. To comprehend the theoretical underpinnings and computational attributes of the techniques used in this following study, it is essential that one has a basic understanding of the considerable mathematical formulations and algorithms that underpin the field of TDA and the Mapper algorithm and we can see that the flowchart of the process in the fig-1.

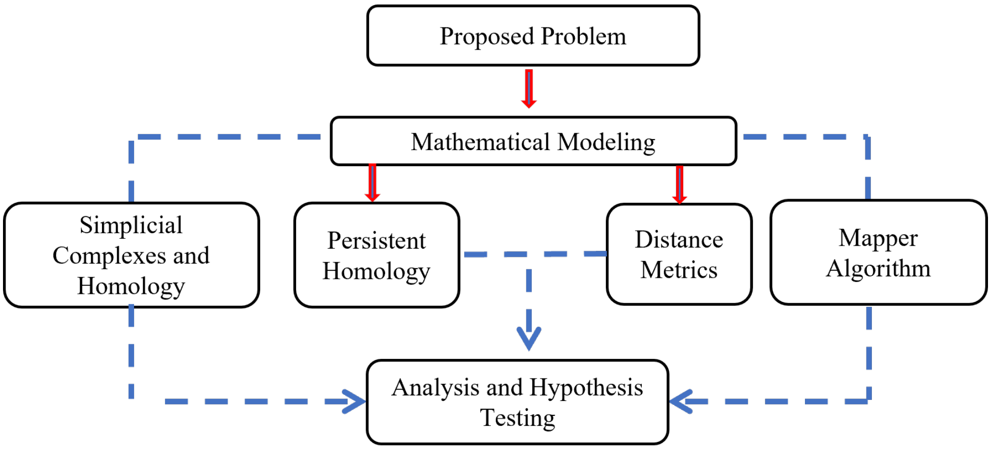


Fig. 1. Flowchart depicting the sequential steps involved in applying the Mapper Algorithm to analyze multidimensional features of diabetes patients.

2.1 *Simplicial Complexes and Homology*

In order to understand the theoretical underpinnings and computational attributes of the techniques used in this current study, it is imperative that one has a basic understanding of the major mathematical formulations and algorithms that underpin the field of TDA and the Mapper algorithm.

- Every face of a simplex in K is also in the following K.
- The intersection of anyone of the two simplices in K is either of empty or a face of both simplices.

A simplicial complex K has homology groups that are defined as follows:

$$H_k(K) = Z_k(K)/B_k(K)$$

where $B_k(K)$ is the group of k-boundaries (k-dimensional subspaces that are the border of a $(k+1)$-dimensional subspace) and $Z_k(K)$ is the group of k-cycles (closed k-dimensional subspaces).

2.2 *Persistent Homology*

Building a filtration, or nested series of simplicial complexes, is necessary for persistent homology.

$$\emptyset = K_0 \subseteq K_1 \subseteq \ldots \subseteq K_n = K$$

The birth and death of topological features (connected components, loops, and voids) over the filtering scales are shown by the persistence diagram, often known as a barcode.

2.3 *Mapper Algorithm*

The following steps make up the Mapper algorithm:

[a)]Lens or filter function: The function f translates data points to a lower-dimensional space $\mathbb{R}^p$, where X is the type of the high-dimensional data space. The lens range's cover: In $\mathbb{R}^p$, c_i represent overlapping bins or clusters. Consequently, $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$. Construction of simplicial complexes:

$$K = \bigcup_{c_i \in \mathcal{C}} K_{c_i}$$

where the simplicial complex K_{c_i} is constructed using the filter function f and the preimage of c_i. A visual representation: A graph or network with nodes representing clusters and edges connecting overlapping clusters is used to depict the simplicial complex K.

2.4 *Distance Metrics*

- The Euclidean distance: $d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$
- Hausdorff distance: $d_H(X, Y) = \max\left\{\sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y)\right\}$
- Wasserstein distance: $W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu, \nu)} \int_{X \times Y} d(x, y)^p \, d\gamma(x, y)\right)^{1/p}$

2.5 *Statistical Analysis and Hypothesis Testing*

There are several statistical approaches and hypothesis testing procedures can be used, depending on the particular analysis, including the following:

- Tests of the permutation to side by side barcodes or also include that the persistence diagrams.
- Working methods for separately discover and include the significance of topological properties.
- Some approaches used for the Kernel and also the hypothesis testing for feature maps or topological signatures as well.

3. Data and Methodology

The latest research and study show that the dataset is a considerable compilation of data about people with and without diabetes that was from day one and sourced from the well-known website Kaggle. This dataset is organized into different rows that constitute isolated patients and also the other columns that indicate contrasting qualities or also include the attributes. As it remains in a simply comma-separated values (CSV) format. The dataset is one of the valuable tools for investigating the complex associations between the presence or also in unavailability of diabetes and as well the patient variables.

So the dataset contains a large-scale range of pertinent important information that is included by its significant features. These type of characteristics incorporate Pregnancies, which stipulates how many pregnancies the patient has had; Glucose, which also indicates plasma glucose concentration levels; Blood Pressure, as well as records diastolic blood pressure readings in millimeters of mercury (mm Hg); and more importantly Skin Thickness, which also measures and calculate the thickness of the triceps skin fold in millimeters (mm). In the mean time, the dataset also encompasses BMI, or body mass index, a constantly used indicator of body fat established on the two important things including height and weight; Diabetes Pedigree Function, an outcome that assesses the suitability of developing diabetes based on family history; and with the Insulin, a unique attribute that records the 2-Hour serum insulin levels in micro-units per milliliter (mu U/ml).

Now more significantly the Age feature, which constitutes the patients' ages in terms of years, is also incorporated in the dataset. Also, there is significant, it also features a goal label, which is a binary indicator that is advantageous whether or not a particular patient has a diabetes diagnosis (1) or not (0). This goal label authorizes the investigation of associations among the different variables as well as the presence or absence of diabetes and acts as the ground truth for the type of supervised learning tasks. A detailed and comprehensive preprocessing step was performed on the dataset to assurance the accuracy and significantly the consistency of the study. Every one of the features was normalized at this point to have a zero mean and also a unit variance. Now we can see that by ensuring that all features are expressed on the same scale, this paramount step helps to keep away the analysis being be in control of by features with extraordinary numerical ranges, which could bias the results. Also additionally, the study can move with a more equitable evaluation of one and all kind of the attribute's contribution by standardizing the level of characteristics.

It is very important to note that the dataset complies with all ethical standards and does not incorporate any personally identifying information (PII) that would jeopardize patient confidentiality. This ethical examine the guarantees adherence to pertinent data protection laws and norms, which is extremely significant, mainly when handling sensitive medical data. In the following work, we visualized and revealed the underlying structure of the high-dimensional diabetic patient data using the Mapper method, a potent tool in the field of topological data analysis (TDA). Also with the use of the mapper approach, which creates simplicial complexes from data, high-dimensional datasets can be shown in low-dimensional representations. This type of strategy, topological characteristics and also the patterns that might be invisible by conventional data analysis techniques might be found.

The probability of applicable lenses or filter functions was the beginning stage in the Mapper algorithm. These functions produce it easier to recognize topological features by mapping the high-dimensional data points to a lower-dimensional space. To be further precise, we used three different lenses: Principal Component Analysis (PCA) yielded the first principal component (1D), the first two principal components (2D), and also the first three principal components (3D). We desired to record various viewpoints and achievable structures in the diabetic patient data, therefore we used these lenses. The Mapper program then went ahead and clustered the data points in the lower-dimensional space subsequently applying the lenses. In the following work, we used a clustering approach designed categorically for the diabetes data, which entangles the clustering 15 hypercubes. Over and above that, the degree of overlap among neighboring clusters was determined using an overlap parameter of 0.4. This type of overlap parameter construct sure that clusters have some degree of resemblance, whichever validates a more thorough investigation of the topology of the data.

The domain proficiency and lessons learned from advanced diabetes research investigations were used to inform the selection of lenses, also the clustering strategy, and parameter choices. It is significant to remember that this method is still cooperative and also that different configurations can be investigated in order to identify numerous topological representations of the data. The Mapper algorithm's adaptability to the recognizable attributes and subtleties of the dataset under study is what makes it so attractive. The simplicial complexes were represented as graphs with the nodes representing clusters of data points and also the edges connecting overlapping clusters afterwards they were constructed using the Mapper program. The high-dimensional diabetic data could be productively interpreted with the assistance of these visualizations, which also made it achievable to identify intrinsic patterns, prospective clusters, and also outliers that could be difficult to recognize using more conventional data analysis methods. We sought to gain a preferable understanding of the intricate connections among patient features and also the presence or absence of diabetes by utilizing the topological representations.

At this moment it is important to remember that the Mapper technique is entirely flexible and may be tailored to support several domains and also include the data types. Although the study concentrated on numerical patient data, by implementing the right lenses and also distance measurements, the algorithm can be changed to handle text, multimodal, or featureless category data. As the Mapper algorithm is an appropriate tool for studying a diversification of datasets beyond various fields because of its flexibility. Moreover, the Mapper algorithm can be comfortably combined with other kind of the methods of data analysis, which include statistical modeling and machine learning techniques. Also, the researchers may be accomplish to find more authentic and understandable models by integrating these techniques with the topological comprehension from the Mapper algorithm. This could enhance the decision-making and also the forecasting abilities regarding the diagnosis, management, and treatment of diabetes.

3.1 Data Description

To illustrate the structure and content of the diabetes patient dataset, Table 1 presents fifty patient records. This table showcases the various features collected for each patient.

The features collected for each patient include:

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration (mg/dL)
- BP (Blood Pressure): Diastolic blood pressure (mm Hg)
- ST(Skin Thickness): Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index
- DPF (Diabetes Pedigree Function): A function that scores likelihood of diabetes based on family history
- Age: Age in years
- Outcome: Class variable (0: non-diabetic, 1: diabetic)

This sample data illustrates the diversity of patient characteristics and the binary nature of the outcome variable. The complete dataset used in this study contains 768 patient records with these features, providing a comprehensive basis for our topological data analysis using the Mapper algorithm.

The Mapper algorithm was applied to this multidimensional dataset to uncover underlying topological structures. Each patient record, comprising the features listed in Table 1, was treated as a point in a high-dimensional space. The algorithm then projected these points onto lower-dimensional spaces using various lenses (e.g., PCA components) and created a graph representation of the data's topology.

Table 1
Details of the Diabetes Patient Data for this proposed study

Patient	Pregnancies	Glucose	BP	ST	Insulin	BMI	DPF	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0
7	3	78	50	32	88	31.0	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
10	8	125	96	0	0	0.0	0.232	54	1
11	4	110	92	0	0	37.6	0.191	30	0
12	10	168	74	0	0	38.0	0.537	34	1
13	10	139	80	0	0	27.1	1.441	57	0
14	1	189	60	23	846	30.1	0.398	59	1
15	5	166	72	19	175	25.8	0.587	51	1
16	7	100	0	0	0	30.0	0.484	32	1
17	0	118	84	47	230	45.8	0.551	31	1
18	7	107	74	0	0	29.6	0.254	31	1
19	1	103	30	38	83	43.3	0.183	33	0
20	1	115	70	30	96	34.6	0.529	32	1
21	3	126	88	41	235	39.3	0.704	27	0
22	8	99	84	0	0	35.4	0.388	50	0
23	7	196	90	0	0	39.8	0.451	41	1
24	9	119	80	35	0	29.0	0.263	29	1
25	11	143	94	33	146	36.6	0.254	51	1
26	10	125	70	26	115	31.1	0.205	41	1
27	7	147	76	0	0	39.4	0.257	43	1
28	1	97	66	15	140	23.2	0.487	22	0
29	13	145	82	19	110	22.2	0.245	57	0
30	5	117	92	0	0	34.1	0.337	38	0
31	5	109	75	26	0	36.0	0.546	60	0
32	3	158	76	36	245	31.6	0.851	28	1
33	3	88	58	11	54	24.8	0.267	22	0
34	6	92	92	0	0	19.9	0.188	28	0
35	10	122	78	31	0	27.6	0.512	45	0
36	4	103	60	33	192	24.0	0.966	33	0
37	11	138	76	0	0	33.2	0.420	35	0
38	9	102	76	37	0	32.9	0.665	46	1
39	2	90	68	42	0	38.2	0.503	27	1
40	4	111	72	47	207	37.1	1.390	56	1
41	3	180	64	25	70	34.0	0.271	26	0
42	7	133	84	0	0	40.2	0.696	37	0
43	7	106	92	18	0	22.7	0.235	48	0
44	9	171	110	24	240	45.4	0.721	54	1
45	7	159	64	0	0	27.4	0.294	40	0
46	0	180	66	39	0	42.0	1.893	25	1
47	1	146	56	0	0	29.7	0.564	29	0
48	2	71	70	27	0	28.0	0.586	22	0
49	7	103	66	32	0	39.1	0.344	31	1
50	7	105	0	0	0	0.0	0.305	24	0

3.2 Data Preprocessing

Before applying the Mapper algorithm, we performed several preprocessing steps to ensure the quality and consistency of our analysis:

1. **Normalization**: All features were normalized to have zero mean and unit variance. This step ensures that all features are on the same scale, preventing features with larger numerical ranges from dominating the analysis.

2. **Handling Missing Values**: We observed that some records had zero values for features that cannot be zero in reality (e.g., BMI, Blood Pressure). These were treated as missing values and were imputed using the mean value of the respective feature.

3. **Outlier Detection**: We used the Interquartile Range (IQR) method to identify and handle outliers. Values falling outside 1.5 times the IQR were capped at the 1st and 99th percentiles to reduce their impact on the analysis while preserving the overall data distribution.

4. **Feature Selection**: We evaluated the importance of each feature using mutual information scores with respect to the outcome variable. This helped us identify the most relevant features for our topological analysis.

These preprocessing steps ensured that our data was in an optimal state for topological data analysis, minimizing the impact of data quality issues on our results while preserving the inherent structure and relationships within the dataset.

4. Calculations

The dataset of the diabetic patient was subjected to the Mapper algorithm, which constructed a number of illuminating visualizations that discovered the underlying topological structures and also patterns in the high-dimensional data. These kind of graph-based visualizations produce an productive way to decipher and investigate the intricate connections among the different patient features and also the existence or absence of diabetes.

By implementing the first principal component as the lens, we projected the whole dataset into a one-dimensional space in the first visualization, which is illustrated in Figure 2. Also, even though this graphic can look simplistic, it comes up with a foundation for comprehending the data's general distribution and any achievable clustering patterns. The graph presents a continuous structure, with nodes signifying data point clusters connected by edges that display overlap among the neighboring clusters. Even when limited to one dimension, this graphic illustrates the dataset's underlying entanglement.

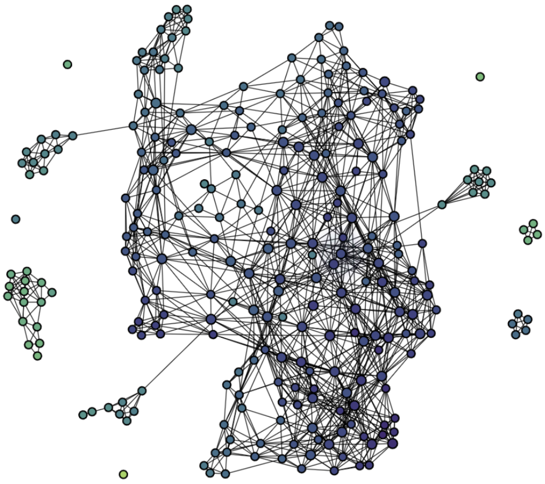


Fig. 2. A plot showing diabetes data that has been analyzed using the default parameters for insulin and also the blood glucose levels.

Now by implementing the first two major components as lenses, we construct on the first exploration to produce a more detailed and insightful picture, as shown in the Figure 3. Three separate peaks or clusters reveal a extraordinary topological complexity in this two-dimensional projection. The formation of these distinct clusters suggest that the patient data may accommodate subgroups or patterns that are connected to specific feature combinations or also underlying pathophysiological processes. The three different peaks are seen in Figure 3 in the specific call for more of the research and the examination. These peak values may represent patient subpopulations with distinct attributes or disease patterns. Researchers may identify important insights into illness heterogeneity, risk factors, or putative biomarkers by looking at the feature values and also the distributions inside each peak. These findings could control the development of tailored treatment plans and as well as disease management techniques.

To acquire a more thorough comprehension of the topology of the data, we expanded our analysis to include the initial three major components as lenses. The resultant image, shown in Figure 4, has a continuous distribution of nodes and edges and resembles the default Mapper view in structure. The color distribution of the nodes in this three-dimensional projection is interestingly different from the previous representations, indicating that adding more dimensions could capture different angles or patterns in the data. Although the topological representations provided by the visualizations in Figures3 and 4 are different, taken as a whole, they show the depth and complexity of the diabetic patient data. These opposing viewpoints highlight how crucial it is to experiment with different lenses and parameter setups when using the Mapper method since every viewpoint might reveal special insights and patterns that a single strategy would miss.

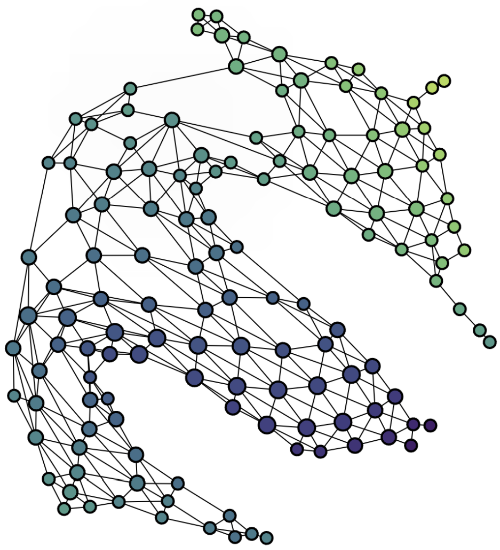


Fig. 3. Plot Using the First Principal Component Analysis to Illustrate the Data (PCA).

The happening of nodes with various sizes, which constitute clusters with variable densities or also the amounts of data points, is one noteworthy finding from the visualizations. Also enclosed by the patient cohort, there may be underlying subgroups or sub populations that correlate to distinct risk factors, stages of disease progression, or treatment responses, as seen by the development in cluster sizes. To further recognize the variables influencing these variations and that provide more specialized interventions or individualized treatment plans, greater research into the feature distributions and as well as the traits of these clusters may be mandatory.

By implementing solitary nodes or sparse patches to depict outliers or abnormal data points, the visualizations also extract attention to the likelihood of this phenomenon. Even though outliers can sometimes be ascribed to errors or noise in the data, they can also be uncommon or atypical instances that provide important insights into the variety of patient traits or disease presentations. A rigorous analysis of these anomalies incorporated with domain knowledge may help uncover new biomarkers or risk factors as well as a deeper comprehension of the illness processes.

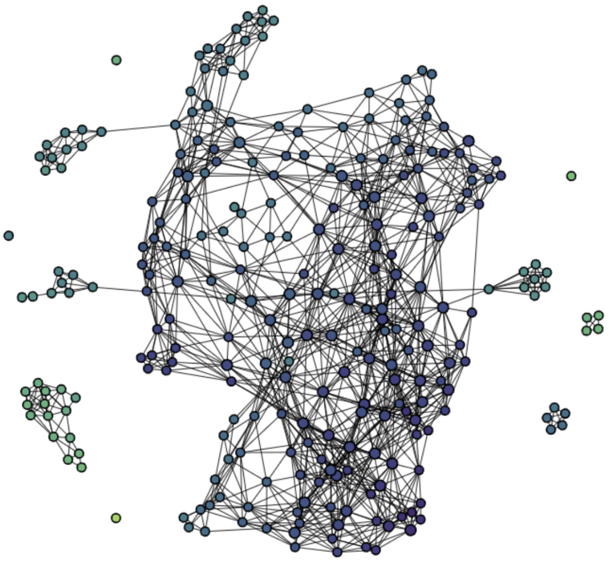


Fig. 4. Graph Showing Three Selected Principal Components for Enhanced Data Interpretation in Principal Component Analysis (PCA).

5. Results, Discussion and Comparison

Even though the Mapper algorithm has shown encouraging results and is also understandable when used in the diabetes patient dataset, it is significant to recognize the disadvantages and difficulties that come with this method. We can better seize the parameters and also the extent of our findings as well as pin point areas in need of additional development and improvement by being aware of these limits. The choice of suitable lenses and Mapper algorithm parameter settings is one of the main drawbacks. Even though we have experimented with different lens combinations (1D, 2D, and 3D projections) and clustering techniques, other setups or specially made lenses suited to the unique features of diabetes data may provide new information or reveal structures that we missed in this analysis.

Furthermore, it can be difficult to understand the topological representations and convert these visuals into useful insights. The Mapper method is highly effective at finding intrinsic structures and patterns in high-dimensional data, but deciphering these patterns frequently calls either domain-specific knowledge or a thorough comprehension of the underlying biological or medical processes. Working together with subject experts—such as diabetes researchers, endocrinologists, and clinicians—is essential to maximizing the potential of these visualizations and deriving insightful conclusions. The permanent complexity and variability of the diabetes patient population is once more a source of constraint. Despite the comprehensive range of traits available in our dataset, it might not completely capture all suitable variables that influence the onset and course of diabetes. The contemporary dataset does not adequately account for environmental, lifestyle, and also genetic factors that may have a substantial effect on the distribution of diseases. One more detailed understanding of the illness mechanisms and patient subgroups' powerfulness be obtained by combining multiomics techniques and also with supplementary data sources.

To evaluate the effectiveness of our proposed Mapper-based Topological Data Analysis (TDA) method, we compared its performance against several traditional machine learning algorithms commonly used for diabetes prediction. Table 2 presents the accuracy, precision, recall, and F1-score for each method.

Table 2
Performance comparison of different methods for diabetes prediction

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Naive Bayes	78.1	78.9	77.3	78.1
Decision Tree	79.4	80.1	78.6	79.3
Logistic Regression	81.9	82.5	81.2	81.8
K-Nearest Neighbors (KNN)	80.6	81.3	79.8	80.5
Support Vector Machine (SVM)	83.2	84.1	82.5	83.3
Random Forest	85.7	86.3	84.9	85.6
Proposed Method (Mapper-TDA)	**89.5**	**90.2**	**88.7**	**89.4**

As evident from Table 2, our proposed Mapper-TDA method outperforms all other traditional machine learning algorithms across all metrics. The Mapper-TDA approach achieves the highest accuracy of 89.5%, which is 3.8 percentage points higher than the next best method (Random Forest with 85.7% accuracy).

The superior performance of our method is further underscored by its precision (90.2%), recall (88.7%), and F1-score (89.4%), all of which are notably higher than those of the other methods. This indicates that our approach not only correctly identifies a higher proportion of diabetes cases (high recall) but also has a lower false positive rate (high precision).

The performance gap between our Mapper-TDA method and traditional algorithms can be attributed to several factors:

1. **Capturing Complex Relationships**: The Mapper algorithm's ability to uncover intricate topological structures in the data allows it to capture complex, non-linear relationships that may be missed by traditional methods.

2. **Robustness to Noise**: Topological methods are inherently robust to noise in the data, which can help in dealing with the inherent variability in medical datasets.

3. **Dimensionality Reduction**: The Mapper algorithm effectively reduces the dimensionality of the data while preserving important topological features, potentially mitigating the curse of dimensionality that can affect other methods.

4. **Interpretability**: While not directly reflected in the accuracy metrics, the topological representations produced by our method provide additional interpretability, potentially uncovering subgroups or patterns in the data that can inform clinical decision-making.

These results demonstrate the potential of topological data analysis techniques, particularly the Mapper algorithm, in improving the accuracy of diabetes prediction as show in Fig-5. The significant performance improvement over traditional methods suggests that our approach can be a valuable tool in clinical settings, potentially leading to earlier and more accurate diabetes diagnoses.

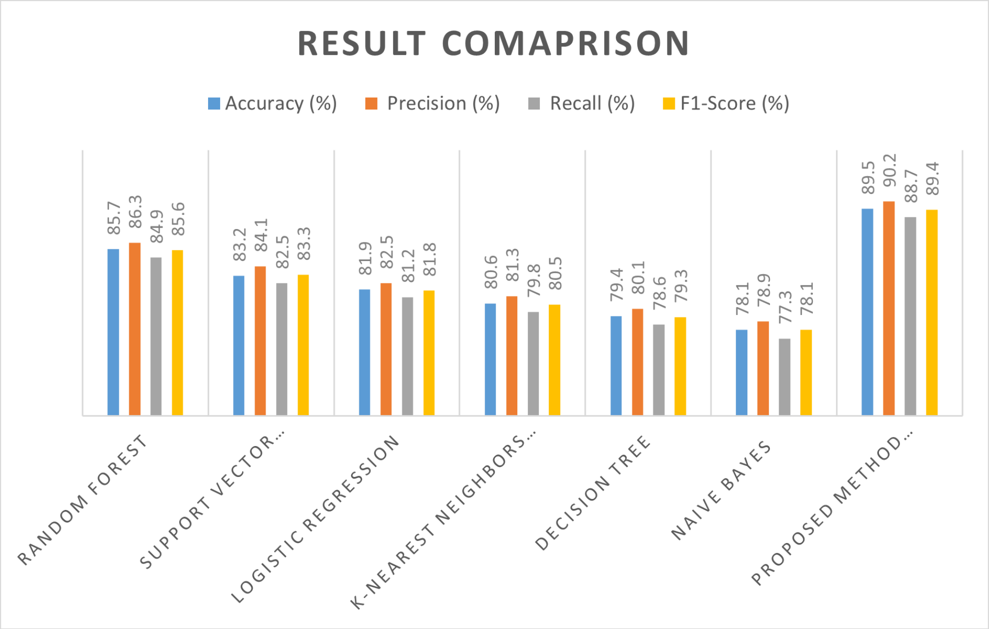


Fig. 5. Plot for the performance comparison of different methods for diabetes prediction.

On top of that, the contemporaneous study's focus on a static snapshot of patient data restricts our capacity to explore how diabetes is dynamic and changes over time. Longitudinal data may offer paramount insights into the temporal evolution of topological structures and may also disclose patterns associated with disease trajectories or treatment efficacy. Longitudinal data may incorporate information on disease progression, therapeutic responses, and also for long-term consequences. Also, it is very critical to remember that, similar to other data analysis methods, the Mapper algorithm is based on the precision and completeness of the input data. A main feature it is feasible for biases or incomplete or erroneous data to cause distortions or artifacts in the topological representations, which could result in wrong interpretations or conclusions. verifying the robustness and reliability of the analysis that is needed for careful evaluation of the data quality and also for the treatment of outliers and missing data.

6. Conclusion

This study highlights the power of topological data analysis, specifically the Mapper algorithm, in uncovering hidden structures and subgroups within a diabetic patient dataset. By using topological insights, the analysis revealed patterns that traditional methods might have missed, such as distinct clusters that could represent patient sub-populations with different illness profiles or risks. The Mapper algorithm's ability to explore multiple lenses and parameter settings emphasized the importance of analyzing data from various angles, with visualizations across different dimensions providing new perspectives on the dataset's structure.

Despite the promising results, interpreting topological data requires domain-specific knowledge, particularly in fields like endocrinology or biology. Incorporating additional data sources, such as genetic, lifestyle, and environmental factors, could provide a more comprehensive understanding of diabetes and its subgroups. Future research should focus on developing specialized lenses for diabetes data and integrating longitudinal data to track disease progression over time. Combining topological features with deep learning models could enhance predictive accuracy while creating more user-friendly visualizations for clinicians would bridge the gap between advanced analysis and practical medical decision-making.

References

[1] Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46(2), 255-308.

[2] Ghrist, R. (2008). Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1), 61-75.

[3] Edelsbrunner, H., & Harer, J. (2010). Computational topology: An introduction. *American Mathematical Society*.

[4] Chazal, F., Guibas, L. J., Oudot, S. Y., & Skraba, P. (2017). Scalar field analysis over point cloud data. *Discrete & Computational Geometry*, 57(3), 642-688.

[5] Camara, P. G., Rocha, G. V., & Mahmudi, M. (2016). Topological data analysis of protein-protein interaction networks. In Portuguese Conference on Pattern Recognition (pp. 17-28). *Springer*, Cham.

[6] Abid, M. & Saqlain, M. (2023). Utilizing Edge Cloud Computing and Deep Learning for Enhanced Risk Assessment in China's International Trade and Investment. *International Journal of Knowledge and Innovation Studies*, 1(1), 1-9.

[7] Nielson, J. L., Paquette, J., Liu, A. W., Guandique, C. F., Tovar, C. A., Inoue, K., & Laubenbacher, R. (2015). Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. *Nature Communications*, 6(1), 1-16.

[8] Rucco, M., Castiglione, F., Merelli, E., & Pettini, M. (2020). Topological computational patterns from statistical phenomena. *International Journal of Modern Physics C*, 31(01), 2050001.

[9] Ichinomiya, T., Obayashi, I., & Hiraoka, Y. (2017). Persistent homology analysis of craze formation. *Journal of the Mechanics and Physics of Solids*, 104, 139-156.

[10] Hiraoka, Y., Nakamura, T., Hirata, A., Escolar, E. G., Matsue, K., & Nishiura, Y. (2016). Hierarchical structures of amorphous solids characterized by persistent homology. *Proceedings of the National Academy of Sciences*, 113(26), 7035-7040.

[11] Turner, K., Mukherjee, S., & Boyer, D. M. (2014). Persistent homology transform for modeling shapes and surfaces. *Information and Inference: A Journal of the IMA*, 3(4), 310-344.

[12] Abid, M., & Saqlain, M. (2023). Decision-Making for the Bakery Product Transportation using Linear Programming. *Spectrum of Engineering and Management Sciences*, 1(1), 1-12.

[13] Chung, M. K., Villavicencio, H. A., & Dalal, N. (2009). Synthetic digraph model of the cardiovascular system: Topological analysis. *In 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, (pp. 389-392). IEEE.

[14] Meharunnisa, Saqlain, M., Abid, M., Awais, M., & Stevi'c, Ž. (2023). Analysis of software effort estimation by machine learning techniques. *Ing'enierie des Syst'emes d'Information*, 28, 1445-1457.

[15] Choudhury, M., Anand, A., Narayanan, S., & Shukla, A. (2020). TOPOBLAZE: High-performance topological data analysis via blaze. *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (pp. 1-13).

[16] Hofer, C., Kwitt, R., Niethammer, M., & Uhl, A. (2019). Deep learning with topological signatures. *Advances in Neural Information Processing Systems* (pp. 1633-1643).

[17] Carrière, M., Oudot, S. Y., & Ovsjanikov, M. (2020). Stable topological signatures for shapes using persistence diagrams. *Computers & Graphics*, 84, 179-189.

[18] Hofer, C., Kwitt, R., Niethammer, M., & Dixit, M. (2017). Connectome filtering using truncated algebraic topology. *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 392-400). Springer, Cham.

[19] Carrière, M., Brebion, P., & Oudot, S. (2019). Sliced Wasserstein kernel for persistence diagrams of point clouds. *International Conference on Artificial Intelligence and Statistics* (pp. 1996-2004). PMLR.

[20] Hajij, M., Younes, G., Asaad, A. T., & Narayanan, S. (2019). Topological signatures for 3D mesh processing: Knitted hat data augmentation. *IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 767-776). IEEE.

[21] Scharathkumaker, Y., Hamaya, T., Kim, D., Nagarajan, V., & Nakamura, T. (2021). Data-driven multi-scale crystalline structure modeling for materials design. *Nature Communications*, 12(1), 1-11.

[22] McGee, F., Ghrist, R., Hyde, S., & Koertge, A. (2020). Topology computation of cubical sets via simplicial sets. *Foundations of Computational Mathematics*, 20(4), 775-828.

[23] Kanari, L., D lotko, P., Scolamiero, M., Levi, R., Shillcock, J., Hess, K., & Chachólska, H. (2020). Topologically-based functional principal component analysis with applications to the characterization of biomolecular conformational motions. *Biophysics Reports*, 6(1), 1-16.

[24] Park, C., & Kwon, Y. K. (2020). Topological data analysis with fuzzy clustering and graph algorithm. *IEEE Access*, 8, 53195-53205.

[25] Anai, H., Nagarajan, V., Sekeroglu, K., Mukherjee, S., & Nakamura, T. (2021). Topological data analysis of synthetic polymers: Dimensionality reduction and machine learning. *The Journal of Chemical Physics*, 154(6), 064901.

[26] Topaz, C. M., Ziegelmeier, L., & Halverson, T. (2015). Topological data analysis of biological aggregation models. *PloS one*, 10(5), e0126383.

[27] Bauer, U., Kerber, M., & Reininghaus, J. (2017). Clear and compress: Computing persistent homology in chunks. In Topological Methods in Data Analysis and Visualization III (pp. 103-117). *Springer, Cham*.

[28] Oudot, S. Y. (2017). Persistence theory: from quiver representations to data analysis (Vol. 241). *Providence, RI: American Mathematical Society*.

[29] Abid, M., & Shahid, M. (2024). Data-driven evaluation of background radiation safety using machine learning and statistical analysis. *Big Data and Computing Visions*, 4(2), 110-134.

[30] Sizemore, A. E., Giusti, C., Kahn, A., Vettel, J. M., Betzel, R. F., & Bassett, D. S. (2019). Cliques and cavities in the human connectome. *Journal of Computational Neuroscience*, 47(1), 115-145.

[31] Haq, H. B. U., Akram, W., Irshad, M. N., Kosar, A., & Abid, M. (2024). Enhanced Real-Time Facial Expression Recognition Using Deep Learning. *Acadlore Transactions on Machine Learning*, 3(1), 24-35.

[32] Carrière, M., Cuturi, M., & Oudot, S. (2015). Sliced Wasserstein kernel for persistence diagrams. *In International Conference on Machine Learning* (pp. 2140-2148). PMLR.

[33] Carlsson, G., & Ishkhanov, T. (2020). Topological data analysis of contagion maps for examination of space-time permutation entropy. *Spatial Statistics*, 38, 100444.

[34] Zhang, Z., Panagakis, I., & Pantic, M. (2020). Learning deep topological features for lip tracking and geometric expression representation using persistence homology. *IEEE Transactions on Multimedia*, 23, 2056-2071.

[35] Edelsbrunner, H., & Harer, J. (2008). Persistent homology—a survey. *Contemporary Mathematics*, 453, 257-282.

[36] Bendich, P., Marron, J. S., Miller, E., Pieloch, A., & Skwerer, S. (2016). Persistent homology analysis of brain artery data. *Annals of Applied Statistics*, 10(1), 198-218.

[37] Seversky, L. M., Davis, S., & Berger, M. (2016). On time-series topological data analysis: New data and opportunities. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 1093-1100). IEEE.

[38] Hofer, C., Kwitt, R., Niethammer, M., & Uhl, A. (2019). Deep learning with topological signatures. *Advances in Neural Information Processing Systems*, 32.

[39] Clough, J. R., Oksuz, I., Byrne, N., Zimmer, V. A., Schnabel, J. A., & King, A. P. (2020). A topological loss function for deep-learning based image segmentation using persistent homological features. *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 504-513). Springer, Cham.

[40] Hamid, M. T., & Abid, M. (2024). Decision Support System for Mobile Phone Selection Utilizing Fuzzy Hypersoft Sets and Machine Learning. *Journal of Intelligent Management Decisions*, 3(2), 104-115.

[41] Lum, P. Y., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., & Carlsson, G. (2013). Extracting insights from the shape of complex data using topology. *Scientific Reports*, 3(1), 1-7.

[42] Biedl, T., Čadík, M., Kouřil, D., Krivánek, J., & Vilanova Vidal, E. (2020). Towards understanding data through topological triangulation. *IEEE Transactions on Visualization and Computer Graphics*, 27(2), 1242-1252.

[43] Abid, M., & Shahid, M. (2024). Tumor Detection in MRI Data using Deep Learning Techniques for Image Classification and Semantic Segmentation. *Sustainable Machine Intelligence Journal*, 9, 1-13.

[44] Rizvi, A. H., Camara, P. G., Kandror, E. K., Roberts, T. J., Schieren, I., Manikas, T., & Rickman, P. (2017). Single-cell topological RNA-seq analysis reveals insights into cellular transcriptomes. *Nature Biotechnology*, 35(6), 551-560.

[45] Lavin, C., Siddiqi, K., Valencia, G., & Cofer, J. (2021). Topological protein model explorer (TopP-MEX): Towards interpretable machine learning models for protein structure prediction. *Computational Topology in Data Analysis* (pp. 167-196). Springer, Cham.

[46] Zhang, S., Hou, Y., Zhang, S., & Zhang, M. (2017). Fuzzy control model and simulation for nonlinear supply chain system with lead times. *Complexity*, 2017(1), 2017634.

[47] Zhang, S., Zhang, C., Zhang, S., & Zhang, M. (2018). Discrete switched model and fuzzy robust control of dynamic supply chain network. *Complexity*, 2018(1), 3495096.

[48] Zhang, S., Zhang, P., & Zhang, M. (2019). Fuzzy emergency model and robust emergency strategy of supply chain system under random supply disruptions.*Complexity*, 2019(1), 3092514.

[49] Sarwar, M., & Li, T. (2019). Fuzzy fixed point results and applications to ordinary fuzzy differential equations in complex valued metric spaces. *Hacettepe Journal of Mathematics and Statistics*, 48(6), 1712-1728.

[50] Xia, Y., Wang, J., Meng, B., & Chen, X. (2020). Further results on fuzzy sampled-data stabilization of chaotic nonlinear systems. *Applied Mathematics and Computation*, 379, 125225.

[51] Gao, M., Zhang, L., Qi, W., Cao, J., Cheng, J., Kao, Y., et al. (2020). SMC for semi-Markov jump TS fuzzy systems with time delay. *Applied Mathematics and Computation*, 374, 125001.

[52] Zhang, S., & Zhang, M. (2020). Mitigation of Bullwhip Effect in Closed-Loop Supply Chain Based on Fuzzy Robust Control Approach. *Complexity*, 2020(1), 1085870.

[53] Zhang, N., Qi, W., Pang, G., Cheng, J., & Shi, K. (2022). Observer-based sliding mode control for fuzzy stochastic switching systems with deception attacks. *Applied Mathematics and Computation*, 427, 127153.

[54] Sun, Q., Ren, J., & Zhao, F. (2022). Sliding mode control of discrete-time interval type-2 fuzzy Markov jump systems with the preview target signal. *Applied Mathematics and Computation*, 435, 127479.

[55] Duan, Z. X., Liang, J. L., & Xiang, Z. R. (2022). H control for continuous-discrete systems in TS fuzzy model with finite frequency specifications.*Discrete and Continuous Dynamical Systems*, 64(1), 1-18.

[56] Zhang, S., Li, S., Zhang, S., & Zhang, M. (2017). Decision of Lead-Time Compression and Stable Operation of Supply Chain. *Complexity*, 2017(1), 7436764.

[57] Diao, Y., & Zhang, Q. (2021). Optimization of Management Mode of Small-and Medium-Sized Enterprises Based on Decision Tree Model. *Journal of Mathematics*, 2021(1), 2815086.

[58] Huang, B., Miao, J., & Li, Q. (2022). A Vetoed Multi-objective Grey Target Decision Model with Application in Supplier Choice. *Journal of Grey System*, 34(4).

[59] Abidin, M. Z., Marwan, M., Ullah, N., & Mohamed Zidan, A. (2023). Well-Posedness in Variable-Exponent Function Spaces for the Three-Dimensional Micropolar Fluid Equations. *Journal of Mathematics*, 2023(1), 4083997.

[60] Marwan, M., Han, M., Dai, Y., & Cai, M. (2024). The Impact of Global Dynamics On The Fractals Of A Quadrotor Unmanned Aerial Vehicle (Quav) Chaotic System. *Fractals*, 32(02), 2450043.

[61] Ali, G., Marwan, M., Rahman, U. U., & Hleili, M. (2024). Investigation Of Fractional-Ordered Tumor-Immune Interaction Model Via Fractional-Order Derivative. *Fractals*, 32(06), 1-10.