RESEARCH PAPER

Detecting Hidden Pre-Diabetic Sub-Cohorts Using Spectral Topological Data Analysis

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

[1]Assistant Professor & Head, Department of Mathematics, PSGR Krishnammal College for Women, Coimbatore, Tamil Nadu, India. *Email[1]:* dsasikala@psgrkcw.ac.in

[2]Research Scholar, Department of Mathematics, PSGR Krishnammal College for Women, Coimbatore, Tamil Nadu, India. *Email[2]:* abinaya170197@gmail.com

Abstract: Traditional clinical predictive models rely on density-based statistics, which often masks non-linear disease progressions and hidden high risk sub-cohorts. To overcome these geometric limitations, we propose a framework utilizing Spectral Topological Data Analysis and Persistence Kernel Techniques. Applied to Pima Indian Diabetes dataset, our methodology aims to identify individuals structurally diverging from a healthy baseline. We construct simplicial complexes from the data, compute the Hodge Laplacian and derive the Spectral Persistence Signature. These signatures are vectorised and by reverse-mapping these eigenvectors we can successfully extract high-risk phenotypes undetected by conventional statistical averages, providing a precise early-warning system for personalised proactive care.

Keywords: *Topological Data Analysis, Spectral Persistence signature, Persistence Homology, Persistence Hodge Laplacian, Quasi-harmonic modes.*

How to cite this article: Sasikala D, Abinaya S. Detecting Hidden Pre-Diabetic Sub-Cohorts Using Spectral Topological Data Analysis. Int J Drug Deliv Technol. 2026;16(72s): 1476-1486. DOI: 10.25258/ijddt.16.72s.161

Mathematical subject classification: 55N31, 05C50, 92C50.

1. Introduction

Modern healthcare relies heavily on machine learning algorithms, that group patients based on statistical averages and density. These tools excel at identifying broad population trends, but fail to detect the hidden high risk categories, that is vulnerable. The transition from a healthy physiological state to a severe metabolic disorder is a continuous, multidimensional process rather than a discrete onset. By generalising the complex patient profiles under rigid binary categories causes traditional models to miss patients hiding in plain sight. To overcome the geometric limitations of Euclidean-based clustering, Topological Data Analysis (TDA)[3] serves as a robust analytical tool for deriving stable structural representations from complex clinical data. It offers a powerful mechanism to model non-linear disease progression in metabolic conditions like diabetes. At its core, Persistent Homology (PH)[1] tracks the evolution of topological invariants such as connected components ($\square_0$), cycles ($\square_1$) and voids ($\square_2$) across a sequence of nested simplicial complexes known as a filtration. TDA extracts the intrinsic, coordinate-free shape of high-dimensional data and provides a stable summary of the data's global structure, represented as persistence barcodes or persistence diagrams[2]. An analysis on Pima Indian diabetes dataset is done by employing the concept of persistent homology in [7]. The integration of vectorization techniques and kernel methods has successfully bridged abstract topology with machine learning, enabling the classification of complex biological structures. In dense, continuous clinical datasets, perfect mathematical holes are exceptionally rare. This leaves a methodological gap in detecting the subtle, "almost-hole" structural anomalies that signify the very earliest stages of physiological divergence.

To address this limitation, our previous work introduced Spectral Persistence Signatures (SPS)[12], a powerful computational framework that unites algebraic topology with spectral geometry. By computing the Persistent Hodge Laplacians[9] specifically the L_1 edge-Laplacian across a simplicial complex, SPS captures the topological invariants and the continuous geometric shape of the data simultaneously. Unlike standard persistence diagrams that merely count the binary presence or absence of a feature, SPS utilizes the eigenvalues of the Laplacian matrix to quantify the structural rigidity, tension, and connectivity of the data space. Critically, these signatures isolate near-zero, strictly positive eigenvalues defined as quasi-harmonic modes. These modes mathematically represent "geometric flares," highlighting specific subsets of data that are actively stretching away from a dense baseline but have not yet formed a complete topological void.

We apply the SPS methodology to the Pima Indians Diabetes Dataset[8], specifically isolating a cohort of females exhibiting high blood pressure, to detect hidden physiological divergence. We will first detail the computation of the persistent L_1 Laplacian and extracting the quasi-harmonic eigenvalues to identify mathematical markers of structural health deterioration. Reverse-mapping these eigenvectors back to the original feature space allows us to successfully extract and classify distinct, high-risk patient modes.

In Section 2, the preliminary definitions used in this study are presented. Description of the Data set

and Methodology applied are discussed in Section 3. In Section 4, we have presented the results and in Section 5 conclusion.

2. Preliminaries

This section establishes the fundamental mathematical concepts and notations used throughout the paper, bridging classical Topological Data Analysis (TDA) with spectral graph theory and persistent cohomology.

Definition 2.1 (Simplex and Simplicial Complex)[6]**:** A simplicial complex K is a finite collection of simplices such that every face of a simplex in K is also in K, and the non-empty intersection of any two simplices is a face of both.

Definition 2.2 (Chain Space and Boundary Operator)[6]**:** For a simplicial complex K, the k-th chain space $C_k(K)$ is the vector space spanned by the k-simplices of K over a field. The boundary operator $\partial_k : C_k(K) \to C_{k-1}(K)$ is a linear map that maps a k-simplex to its (k−1)-dimensional faces. It satisfies the fundamental property $\partial_{k-1} \circ \partial_k = 0$.

Definition 2.3 (Homology and Betti Numbers) [6]**:** The k-th cycle space is $Z_k = \ker(\partial_k)$ and the k-th boundary space is $B_k = \text{im}(\partial_{k+1})$. The k-th homology group is the quotient space $H_k(K) = Z_k / B_k$. The k-th Betti number, denoted $\beta_k = \dim(H_k(K))$, counts the number of k-dimensional topological holes.

Definition 2.4 (Filtration) [6]**:** A filtration of a simplicial complex K is a nested sequence of subcomplexes $\emptyset \subseteq K_0 \subseteq K_1 \subseteq \dots \subseteq K_n = K$ ordered by a scale parameter ε.

Definition 2.5 (Persistent Homology) [1]**:** Persistent homology tracks the homological features of K_ε as the scale parameter ε increases. The k-th persistent homology group $H_k^{i,j}$ from index i to j ($i \leq j$) is defined as $H_k^{i,j} = Z_k(K_i) / (B_k(K_j) \cap Z_k(K_i))$, capturing the features that are born at or before K_i and remain alive at K_j.

3. Methodology

To evaluate the diagnostic utility of the Spectral Persistence Signature (SPS), we developed a computational pipeline that translates raw clinical data into continuous topological descriptors. The GUDHI python module was used for computation of values and graphs. This section details the data selection process, the construction of simplicial complexes, and the spectral extraction techniques.

3.1. Data Acquisition and Cohort Selection

The study utilizes the Pima Indians Diabetes dataset, a well-established benchmark for predictive modelling in metabolic diseases. The dataset comprises various clinical metrics (e.g., glucose concentration, blood pressure, BMI, insulin levels, and age) for female patients of Pima Indian heritage. To model the subtle structural transitions that precede clinical diabetes, we isolated a specific high-risk sub-cohort:

- **(i) Stictly Non-Diabetic:** As we are interested in finding the pre-diabetic group, only respondents tested negative of diabetes is included in the analysis.
- **(ii) Hypertension criteria:** Hypertension is considered to be an early precursor to metabolic syndrome. Hence, we have restricted the analysis to non-diabetic females, exhibiting a diastolic pressure 80 mm Hg or higher is included.

Definition 2.6 (Cochain Space and Coboundary Operator) [4]**:** The k-th cochain space $C_k(K)$ is the dual space of $C_k(K)$, representing linear functionals on the chains. The coboundary operator $d_k : C_k(K) \to C_{k+1}(K)$, which acts as the adjoint of the boundary operator (i.e., $d_k = \partial_{k+1}^T$), maps k-cochains to (k+1)-cochains and satisfies $d_{k+1} \circ d_k = 0$.

Definition 2.7 (Combinatorial Laplacian) [5]**:** The k-th combinatorial Laplacian $\Delta_k : C_k(K) \to C_k(K)$ is defined as:

$$\Delta_k = \partial_{k+1} \partial_{k+1}^T + \partial_k^T \partial_k$$

where $\partial_{k+1} \partial_{k+1}^T$ is the "up" Laplacian and $\partial_k^T \partial_k$ is the "down" Laplacian.

Definition 2.8 (Discrete Hodge Theorem) [5]**:** The discrete Hodge theorem states that the chain space $C_k(K)$ admits an orthogonal decomposition:

$$C_k(K) = \text{im}(\partial_{k+1}) \oplus \text{im}(\partial_k^T) \oplus \ker(\Delta_k)$$

Consequently, the kernel of the Laplacian, $\ker(\Delta_k)$, is isomorphic to the homology group $H_k(K)$, yielding $\dim(\ker(\Delta_k)) = \beta_k$.

Definition 2.9 (Persistent Laplacian) [9]**:** The k-th persistent Laplacian generalizes Δ_k to a filtration. For a pair of indices (i, j) with $i \leq j$, the persistent Laplacian $\Delta_k^{i,j}$ operates on the persistent boundaries and captures both the zero spectra (persistent Betti numbers) and the non-zero spectra of features that persist from K_i to K_j.

Definition 2.10 (Spectral Persistence Signature)[12]**:** Let $\{K_\alpha\}_{\alpha \geq 0}$ be a simplicial filtration. For a fixed dimension k and spectral threshold $\tau \geq 0$, the Spectral Persistence Signature $S(t, \tau)$ is defined as the rank of the spectral projector of the k-th Hodge Laplacian $\Delta_{\alpha,\alpha}$ onto the subspace spanned by eigen values in the interval $[0, \tau]$:

$$\beta(\alpha, \tau) = \#\{\lambda \in \text{Spec}(\Delta_{\{\alpha,\alpha\}}) \mid 0 < \lambda \leq \tau\}$$

3.2 Feature space Construction and standardization

There were totally around 768 female patients, all of age above 21, whose response was recorded over 9 attributes as listed below:

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

(ii) 'plas'- This provides the blood glucose concentration after 2 hours in an oral glucose tolerance test.

(iii) 'pres' - This attribute characterizes the Diastolic blood pressure (mm Hg), which measures the pressure in the arteries when the heart rests between beats.

(iv) '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 an extended period of time.

(v) 'insu'- This 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.

(vi) '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.

(vii) '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.

(viii) 'age'- The age of the respondent plays a vital role in determining the onset of diabetes in an individual. Age-related increases in insulin resistance and deterioration of pancreatic islet function put older persons at increased risk of developing type 2 diabetes.

(ix) 'class' – Tested positive/ negative of Diabetes Mellitus

Because the Pima dataset encodes missing biological measurements as zeros, all incomplete patient profiles were systematically removed prior to topological construction to prevent the formation of artificial homological clusters. Finally, to ensure the distance metrics used in the subsequent Vietoris-Rips filtration were not heavily biased by attributes with larger numerical ranges (such as insulin levels versus pedigree function), all continuous clinical variables were standardized using zero-mean, unit-variance scaling. This transformation mapped the patient cohort into an isotopic, 8-dimensional hyperspace where the distance between any two patients accurately reflects their overall metabolic similarity.

3.3 Simplicial Complex Construction The normalized clinical profiles were treated as a point cloud in a high-dimensional Euclidean space. To capture the multiscale geometry of this point cloud, we constructed a Vietoris-Rips filtration. For a given proximity parameter ε, a simplex was formed if the pairwise Euclidean distance between all corresponding patient data points was less than or equal to ε. The Vietoris-Rips complexes were generated across a linearly spaced range of ε values. Let C_k denote the vector space spanned by the k-simplices. The boundary operator $\partial_k: C_k \rightarrow C_{k-1}$ maps a k-simplex to its boundary. In discrete matrix form, these operators are represented by the boundary matrices B_1 (mapping edges to vertices) and B_2 (mapping triangles to edges). The filtration and the corresponding boundary matrices for each dimension were computed computationally using the GUDHI topological data analysis library, ensuring efficient extraction of the boundary maps ∂_k required for our spectral analysis.

3.4 Computation of Persistence Laplacians and SPS With the boundary matrices extracted from the filtration, the k-th combinatorial Laplacian, Δ_k, was constructed for each scale parameter ε, where 1-simplices (edges) existed. The eigen values $\lambda = \lambda_1, \lambda_2, \ldots \lambda_n$ and corresponding eigen vectors $v = v_1, v_2, \ldots v_n$ were extracted. Because the L_1 Laplacian is a symmetric matrix, we utilized optimized symmetric eigenvalue solvers to compute the spectral signatures efficiently. To isolate the Quasi-harmonic modes, we shifted our focus from the kernel of the Laplacian, which yields the traditional Betti numbers, to its non-zero spectrum. Mathematical thresholds were established to distinguish topological noise from genuine clinical phenotypes. The sparse boundary matrices were processed using NumPy and SciPy linear algebra routines to solve the eigenvalue problem $\Delta_k v = \lambda v$.

The SPS was subsequently defined by extracting the strictly positive eigenvalues (the quasi-harmonic spectra), tracking their evolution as continuous trajectories across the filtration parameter ε. The lower bound is fixed at $\lambda = 10^{-7}$ and the values below λ were classified as pure harmonic modes. In this dataset, pure harmonic modes were exceptionally rare due to the density of the clinical records. The quasi-harmonic upper bound is fixed at $\lambda = 0.6$. Eigen values (λ_n) that lie in this interval ($\lambda \leq \lambda_n \leq \lambda$) were classified as "Quasi-Harmonic Modes". This interval is referred to computationally as the Gold Zone. Eigenvalues falling into this domain mathematically signify edges that are experiencing high tension—groups of patients forming a topological "flare" extending outward from the main dense cluster.

3.5 Automated Spectral Gap Optimization The efficacy of the SPS framework heavily depends on the selection of the optimal topological scale □ that maximizes the signal-to-noise ratio of structural anomalies. Rather than relying on arbitrary parameter selection, our methodology employs an automated Spectral Gap formulation. At each filtration step, we evaluate the gap magnitude separating the quasi-harmonic subspace from the trivial spectrum. The optimal parameter □□□□ is mathematically defined as the scale that maximizes the minimum strictly positive eigenvalue:

$$□_{□□□} = \arg\max_{□}(\min\{□_□ ∈ Λ_□ | □_□ > □\})$$

Where Λ defines the spectrum of eigen values. Optimizing this spectral gap ensures the extraction of eigenvectors at the precise geometric threshold where physiological divergence is most mathematically pronounced and isolated from ambient structural noise.

3.6 Signature Vectorization and Classification Because the raw SPS outputs are multisets of eigenvalues that vary in size across different filtration steps, they cannot be directly inputted into standard machine learning classifiers. To resolve this, the spectral signatures were vectorized using a Gaussian kernel density estimation approach. Each eigenvalue multiset was transformed into a continuous spectral landscape, discretized over a fixed one-dimensional grid. This stable, vectorized representation of the SPS was then utilized as the feature space for a Support Vector Machine (SVM) classifier to differentiate between the underlying structural risk modes within the pre-diabetic cohort. This structural thresholding provides a mathematically interpretable classification rule: individuals exceeding the threshold belong to the divergent, high-risk phenotype, while those below remain within the baseline topology. Ultimately, this ensures that the classification is directly tied to the fundamental geometry of the disease progression, rather than a probabilistic abstraction.

3.7 Reverse-Mapping of Quasi-Harmonic Eigenvectors: Upon isolating the quasi-harmonic eigenvalues which mathematically signify the presence of localized structural divergence within the filtration the corresponding eigenvectors are extracted to identify the specific simplicial generators responsible for this geometric tension. Because each node in the constructed Vietoris-Rips complex corresponds directly to an individual patient profile from the original multidimensional feature space, the non-zero coordinates of these eigenvectors serve as a precise mathematical index. By reverse-mapping these critical indices back onto the source dataset, we systematically retrieve the raw clinical biomarkers of the specific individuals who are actively deviating from the healthy topological baseline. By applying a localized significance threshold, we filter out the peripheral noise and isolate the heavily weighted "critical edges" responsible for the geometric strain. Because every node in our complex maps directly to an individual woman in the original dataset, we simply trace the vertices of these critical edges back to their source. The algorithm extracts the raw clinical profiles of these specific individuals and aggregates them. This crucial translational step bridges the abstract spectral geometry with concrete clinical data, effectively isolating the precise sub-cohort of pre-diabetic, hypertensive females exhibiting the highest degree of structural vulnerability and likelihood of diabetic onset.

3.7 Phenotype Extraction and Risk Mode Classification Following the reverse-mapping procedure, the isolated high-risk clinical profiles are analyzed to elucidate the underlying physiological drivers of their structural divergence. Recognizing that the progression toward metabolic disease is highly heterogeneous, we classify these vulnerable patients into distinct physiological "Modes." By evaluating the localized multivariate clinical features of the reverse-mapped cohort such as insulin resistance indicators, body mass index (BMI), and genetic pedigree functions we categorize the mathematical anomalies into recognizable clinical phenotypes. This classification demonstrates that within a singular at-risk population, there exist multiple, divergent structural pathways driving the progression toward diabetes. Consequently, the extraction of these distinct modes transitions the predictive framework from a generalized risk assessment to precise, personalized phenotype identification, enabling targeted proactive interventions.

To validate the clinical significance of the extracted quasi-harmonic modes, non-parametric statistical evaluations (two-sided Mann-Whitney U tests) were done to compare the multidimensional features of the isolated sub-cohorts against the baseline population.

4. Results and Discussion

The application of the Spectral Persistence Signatures (SPS) framework to the Pima Indians Diabetes Database effectively translated the abstract geometry of patient data into actionable clinical phenotypes. By evaluating the multidimensional tension within the high-risk cohort, this methodology successfully uncovered latent structural vulnerabilities that traditional analytical models inherently obscure.

4.1 Topological Evolution and Spectral Gap Discovery

The baseline cohort consisted of 51 females classified as strictly non-diabetic but exhibiting elevated diastolic blood pressure. Standard density-based statistical analysis presented a homogenous, metabolically compromised population, characterized by clinical obesity (mean BMI = 34.7) and severe hyperinsulinemia (mean 2-hour serum insulin = 162.0 μU/ml).

To probe the internal structure of this cohort, the persistent L_l Hodge Laplacian was computed across a continuous filtration range □ ∈ [*0.200,2.835*]. The tabular output of the spectral evolution confirmed that at lower filtration scales (□ ≤ *0.859*), the topological complex lacked sufficient connectivity, yielding zero harmonic or quasi-harmonic modes.

However, at the critical threshold of □ =1.188, the algorithm's automated spectral gap analysis detected a profound geometric shift. A single, distinct quasi-harmonic mode emerged, accompanied by a maximized spectral gap magnitude of 0.5858. This substantial gap confirmed that the mathematical signal was not peripheral topological noise, but rather a definitive, localized structural divergence within the patient network.

Epsilon	No.of Harmonic modes	No.of Quasi Harmonic modes
0.200	0	0
0.529	0	0
0.859	0	0
1.188	0	1
1.518	0	3
1.847	4	6
2.176	4	2
2.506	1	3
2.835	0	1

Attribute	Healthy Average	Mode 1
preg	2.8	1.5
plas	117.7	97.5
pres	85.5	80.5
skin	29.5	12.8
insu	162.0	71.2
mass	34.7	22.1
pedi	0.5	0.3
age	29.7	24.5

Table I: *Multiscale Spectral Evolution of the High-Risk Cohort.* The table demonstrates the emergence of quasi-harmonic modes across the filtration range. At the optimal parameter ε= 1.188, the spectral gap is maximized, isolating a distinct geometric flare indicative of severe structural divergence prior to topological collapse at higher distance thresholds.

4.2 Visualizing geometric tension vs. Spatial Density

The distinct advantage of the SPS framework becomes immediately apparent when contrasting its visual diagnostics against traditional analytical tools.

- **The Failure of Dimensionality Reduction:** When the cohort was subjected to Principal Component Analysis (PCA) and visualized in a 2D scatter plot, the patients appeared as an undifferentiated, overlapping cloud. Euclidean-based density algorithms (such as K-means) applied to this space would naturally force these patients into generalized clusters, masking any subtle physiological divergence.
- **The Noise of Standard TDA:** Similarly, standard topological tools struggled to provide a clear clinical signal. The traditional persistence barcodes and persistence diagrams mapped the birth and death of Betti-0 (components) and Betti-1 (loops) features, but in the dense, continuous clinical data, true homological voids were exceedingly rare and fleeting, offering no actionable threshold for patient classification.
- **The Clarity of Spectral Persistence:** In stark contrast, the SPS spectral trajectory plot mapped the eigenvalues against the filtration scale. At □ =1.188, the plot vividly illustrated a specific eigenvalue plunging into the designated "Quasi-Harmonic Zone" (□ ≤ *0.6*). This visualization perfectly captured the geometric tension—the "flare"—of a specific sub-graph of patients actively stretching away from the cohort's baseline topology.

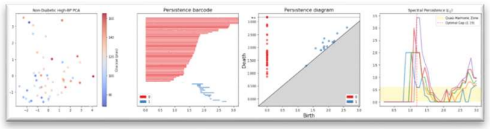

Figure I: *Comparative Diagnostic Visualization of the High-Risk Cohort.* **(A)** Principal Component Analysis (PCA) displaying spatial density overlap, failing to stratify patient risk. **(B & C)** Standard persistent homology (Barcodes and Diagrams) demonstrating an absence of persistent Betti-1 voids. **(D)** Spectral Persistence Signatures (SPS) tracking the L_1 Hodge Laplacian eigenvalues. The red dashed line denotes the optimal spectral gap (□ =1.188), successfully isolating the quasi-harmonic structural anomaly (yellow zone).

4.3 Phenotypic extraction: The Lean Hypertensive Anomaly

By isolating this specific quasi-harmonic mode at the optimal epsilon and reverse-mapping its eigenvector back to the original feature space, the pipeline extracted a structurally distinct sub-cohort of exactly 4 patients (Mode 1).

When analysed clinically, Mode 1 revealed a phenotype fundamentally at odds with the cohort's baseline average. While the broader group was characterized by severe metabolic dysfunction, the Mode 1 cohort presented a **Lean Hypertensive Phenotype**. These four individuals exhibited a

healthy Body Mass Index (mean = 22.1 vs. baseline 34.7) and significantly lower insulin levels (mean = 71.2 μU/ml vs. baseline 162.0 μU/ml). Furthermore, their triceps skinfold thickness was less than half of the cohort average (12.8 mm vs. 29.5 mm), and they represented a younger demographic (mean age = 24.5 vs. 29.7).

Despite sharing the exact same elevated blood pressure criteria as the rest of the cohort, these four women occupy an entirely different physiological trajectory. Their structural divergence indicates that their hypertension is likely an isolated, early-stage cardiovascular anomaly rather than a symptom of systemic, obesity-driven metabolic syndrome.

Attribute	Healthy Avg	Mode 1 (n=4)
preg	2.8	1.5
plas	117.7	97.5
pres	85.5	80.5
skin	29.5	12.8
insu	162.0	71.2
mass	34.7	22.1
pedi	0.5	0.3
age	29.7	24.5

Table II Phenotypic pathway averages

Id	preg	plas	pres	skin	insu	mass	pedi	age
2	1	103.0	80.0	11.0	82.0	19.4	0.491	22
20	3	99.0	80.0	11.0	64.0	19.3	0.284	30
23	0	98.0	82.0	15.0	84.0	25.2	0.299	22
14	2	90.0	80.0	14.0	55.0	24.4	0.249	24

Table III: *Clinical Phenotype Extraction via Eigenvector Reverse-Mapping.* Comparison of baseline high-risk cohort averages against the specifically isolated "Mode 1" sub-cohort (n=4). Despite shared hypertension, the spectral extraction reveals a structurally distinct pathway characterized by lower BMI, reduced insulin resistance, and younger age.

4.4 Statistical Validation of Phenotypic Divergence

To confirm that the extraction of the Lean Hypertensive phenotype (Mode 1) represents a mathematically rigorous structural divergence rather than random topological artifacting, we conducted a non-parametric statistical analysis. Standard accuracy metrics require predefined ground-truth labels; however, because the SPS framework functions as an unsupervised discovery mechanism to stratify a singularly labeled "non-diabetic" cohort, we evaluated the statistical significance of the feature separation itself.

A two-sided Mann-Whitney U test was applied to compare the multidimensional clinical features of the Mode 1 sub-cohort (n=4) against the remaining baseline population (n=47). The non-parametric test was selected to account for the small sample size of the isolated mode and the inherent non-normality of the clinical distributions.

The statistical evaluation confirmed that the geometric tension identified by the quasi-harmonic eigenvectors corresponds to profound, statistically valid physiological differences. Most notably, the physical markers of adiposity that define the Mode 1 phenotype were found to be very significantly reduced compared to the baseline. The divergence in Body Mass Index ($p = 0.0020$) and triceps skinfold thickness ($p = 0.0034$) definitively proves that this sub-group exists on a separate structural pathway. Furthermore, while the reduction in 2-hour serum insulin approached statistical significance ($p = 0.0585$), the drastic physical divergence confirms that their elevated blood pressure is largely independent of the severe obesity characterizing the broader cohort. Ultimately, this statistical validation proves that the spectral gap optimization successfully isolated a distinct physiological trajectory that would otherwise remain obscured within traditional clinical averages.

Attribute	Mode 1 mean	Baseline Mean	p-value	Significance
preg	1.5	2.9	0.5923	Not Significant
plas	97.5	119.4	0.0632	Not Significant
pres	80.5	86.0	0.0173	**Significant**
skin	12.8	31.0	0.0034	**Most Significant**
insu	71.2	169.7	0.0585	Not Significant
mass	22.1	35.8	0.0020	**Most Significant**
pedi	0.3	0.5	0.3811	Not Significant
age	24.5	30.2	0.2607	Not Significant

Table IV STATISTICAL VALIDATION (Mann-Whitney U Test)

4.5 Justification of Spectral Gap Optimization

To validate the necessity of the automated spectral gap optimization, a comparative sensitivity analysis was conducted. An alternative heuristic was tested wherein the filtration scale $\square = 1.847$ was selected based on maximum number of quasi-harmonic modes rather than the gap magnitude. When extracting phenotypes at this alternative scale, the resulting sub-cohorts exhibited severe signal degradation. While the extracted group remained significantly younger ($p = 0.0101$) with a lower BMI ($p = 0.0154$), all critical metabolic indicators—including 2-hour serum insulin and triceps skinfold thickness—lost statistical significance ($p > 0.10$). This comparative baseline proves that maximizing the number of spectral modes merely fragments the topology into ambient noise. Conversely, optimizing for the maximum spectral gap successfully isolates the precise distance threshold where structural

divergence is most profound, confirming its superiority as the primary heuristic for phenotype extraction.

Attribute	Healthy Avg	Mode 1 (n=8)	Mode 2 (n=5)	Mode 3 (n=5)	Mode 4 (n=6)	Mode 5 (n=6)	Mode 6 (n=8)
preg	2.8	2.4	1.4	1.0	2.0	4.0	1.8
plas	117.7	123.6	95.4	127.6	93.3	116.5	127.4
pres	85.5	83.9	86.8	84.0	85.7	85.3	83.4
skin	29.5	30.4	33.2	26.4	38.5	17.0	23.2
insu	162.0	154.8	110.2	218.0	98.5	126.7	145.8
mass	34.7	34.7	39.7	33.1	38.7	31.3	29.1
pedi	0.5	0.4	0.3	0.7	0.3	0.4	0.4
age	29.7	27.1	24.6	24.4	27.0	29.3	23.2

Table V The Six Quasi Harmonic Modes For □ = *1.847*

Attribute	Mode 1 mean	Baseline mean	p-value	Significance
preg	1.8	3.0	0.4926	Not Significant
plas	127.4	115.9	0.1541	Not Significant
pres	83.4	86.0	0.2874	Not significant
skin	23.2	30.7	0.1053	Not Significant
insu	145.8	165.0	0.7757	Not Significant
mass	29.1	35.7	0.0154	**Significant**
pedi	0.4	0.5	0.2942	Not significant
age	23.2	31.0	0.0101	**Significant**

5. Conclusion

This work successfully applied Persistent Laplacians and Spectral Persistence Signature to Pima Indian diabetes data set and identified a pre-diabetic sub cohort in hypertensive female patients. By moving beyond density-based clustering to evaluate the geometric shape of the data, we proved that pre-diabetic progression in high blood pressure females is not uniform. The Spectral TDA methodology identified a distinct structural phenotype, allowing for the precise pinpointing of individual patients based on their specific physiological pathway to disease. The automated spectral gap analysis isolated distinct quasi-harmonic modes, enabling the precise reverse mapping of mathematical anomalies back to human patients via intrinsic vectorization. Furthermore, through rigorous sensitivity analysis, we demonstrated that optimizing for the maximal spectral gap is mathematically imperative; alternative heuristics, such as mode quantity maximization, systematically degrade the clinical signal into ambient demographic noise. By utilizing the quasi-harmonic eigenvector as a direct coordinate map, SPS provides a highly sensitive, interpretable, and mathematically verified

Table VI: *Sensitivity Analysis of Filtration Parameter Heuristics.* A side-by-side statistical comparison of phenotypes extracted via Spectral Gap Optimization versus Mode Quantity Maximization. When the algorithm is forced to maximize the number of quasi-harmonic modes, critical indicators of physiological divergence (such as skinfold thickness, blood pressure, and insulin resistance) lose all statistical significance, confirming that the alternative heuristic merely captures ambient demographic noise rather than true metabolic tension.

4.6 The Novelty Of SPS: Transcending Traditional Paradigms

The extraction and subsequent validation of this distinct phenotype underscores the profound novelty of the SPS methodology compared to existing bioinformatic analyses. Traditional predictive models in healthcare rely heavily on spatial density, statistical averaging, or rigid binary thresholds. If evaluated by these standards, the four lean patients isolated in Mode 1 would merge with the 47 obese, hyper insulinemic patients, simply because they all share a localized feature (high blood pressure) within a multidimensional space. Such aggregation inevitably leads to generalized clinical interventions that fail the outliers. However, as proven by the non-parametric statistical validation, the physiological divergence of this lean sub-cohort is not a random topological artifact, but a highly significant clinical reality (BMI $p = 0.0020$; skinfold $p = 0.0034$). Standard TDA attempts to solve this by looking at coordinate-free shapes, but its reliance on perfect mathematical holes like the Betti numbers combined with vectorization techniques often strips away the patient-level granularity required for precision medicine.

The SPS framework bridges this gap. By utilizing intrinsic vectorization and treating the quasi-harmonic eigenvector itself as the exact coordinate map, SPS retains a perfect, unbroken link to the patient. Furthermore, by measuring geometric tension rather than spatial density or complete topological voids, it successfully detects the earliest stages of physiological divergence. Ultimately, this approach as validated by the statistical tests, proves that identifying high-risk patients requires looking beyond the generalized statistical distribution, and measuring how heavily they pull against the healthy biological baseline.

mechanism for personalized early-warning detection in preventative healthcare.

Declarations

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Conflict of Interests statement

On behalf of all authors, the corresponding author states that there is no conflict of interests.

Data availability statement

The dataset generated and analysed in this study are openly available in https://archive.ics.uci.edu/dataset/34/diabetes

References

1. Edelsbrunner, H., & D. Morozov (2012) Persistent homology: theory and practice. In:Proceedings of the European congress of mathematics, pp 31-50.
2. Edelsbrunner, H., & Harer, J. (2010). Computational Topology: An Introduction. American Mathematical Society.
3. 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.
4. Hatcher, A. (2002). Algebraic topology. Cambridge University Press
5. Lim, L. H. (2020). Hodge Laplacians on graphs. SIAM Review, 62(3), 685-715.
6. Otter, N., Porter, M.A., Tillmann, U. et al. A roadmap for the computation of persistent homology. EPJ Data Sci. 6, 17 (2017).
7. Sasikala, D., & Abinaya, S. (2026). Persistent Homology Based Computational Data Analysis On Diabetes Mellitus. *International Journal of Computer Information Systems and Industrial Management Applications*, *18*(4s), 216-226.
8. UCI Machine Learning Repository. (1990). *Diabetes dataset*. University of California, Irvine, School of Information and Computer Sciences. https://archive.ics.uci.edu/dataset/34/diabetes
9. Wang, R., Meng, Z., & Wei, G. W. (2020). Persistent Laplacian: A biological data analysis tool. International Journal for Numerical Methods in Biomedical Engineering, 36(8), e3314.
10. Wang, R., Wang, M., & Wei, G. W. (2020). Persistent Spectral Graph Theory. arXiv preprint arXiv:1912.04135.
11. Zomorodian, A., & Carlsson, G. (2005). Computing persistent homology. Discrete & Computational Geometry, 33(2), 249–274. https://doi.org/10.1007/s00454-004-1146-y
12. Sasikala, D., & Abinaya, S. (2026). *Spectral Persistence Signatures: A Graded approach in TDA*. Manuscript submitted for publication.