TDA based Feature Extraction for Physiochemical Quality assessment: A Machine Learning Framework with Benchmark Validation

Dr. D. Sasikala[1], Mrs. M. Renukadevi[2]
[1] Assistant Professor & Head, Department of Mathematics, PSGR Krishnammal College for Women, Coimbatore-641004, Tamil Nadu, India. E-mail: dsasikala@psgrkcw.ac.in
[2] Assistant Professor, Department of Mathematics, KPR College of Arts Science and Research, Coimbatore-641407, Tamil Nadu, India. E-mail: renukadevi.velumani@gmail.com

Abstract
Quality assessment based on physiochemical characteristics plays a significant role in food processing, pharmaceutical manufacturing, and fermentation-based product development. The original feature space, which may not fully capture the inherent structural relationships seen in complicated datasets, is typically the basis for conventional machine learning techniques. Recently, topological data analysis (TDA) has become a potent mathematical framework that can extract strong topological and geometric properties from high-dimensional data. Using the Wine Quality dataset as a standard physicochemical dataset, this paper suggests a TDA-based feature extraction framework for multi-class quality classification. The dataset is first preprocessed using missing-value verification and normalization. Then, topological descriptors that describe the underlying data manifold are extracted via persistent homology. An enriched feature representation is created by combining these descriptors with the initial physiochemical characteristics. For quality classification, three supervised machine learning classifiers are used: Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbor (KNN). When compared to traditional feature representations, experimental evaluation shows that the inclusion of topological features enhances classification performance. The suggested framework offers a potential method for quality assessment in fermentation processes, pharmaceutical manufacturing, and other chemical quality evaluation applications. It also demonstrates how TDA may extract buried structure information from physicochemical datasets.

Keywords: Topological Data Analysis, Persistent Homology, Multi-Class classification, Wine Quality, Feature Extraction, Chemometrics.
How to cite this article: Sasikala D, Renukadevi M. TDA Based Feature Extraction for Physiochemical Quality Assessment: A Machine Learning Framework with Benchmark Validation. Int J Drug Deliv Technol. 2026;16(72s): 1487-1493. DOI: 10.25258/ijddt.16.72s.162

1. Introduction

A common issue in food science, agrochemistry, and pharmaceutical formulation research is the chemical characterisation of consumable goods and the estimation of their quality grade from quantifiable physicochemical criteria. Because it combines an ordinal, expert-assigned quality score with a moderate number of continuous, correlated physicochemical descriptors (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free and total sulfur dioxide, density, pH, sulphates, and alcohol content), wine quality classification is a commonly studied benchmark for this class of problems [1]. Since wine samples are usually graded on a discrete scale rather than a binary accept/reject decision, the problem is readily structured as multi-class classification.

Supervised learning methods, such as Random Forests [9], Support Vector Machines [10], and k-Nearest Neighbors trained directly on the raw physicochemical measurements, are used in traditional approaches to this problem. When it comes to minority classes at the extremities of the quality scale, where the sample density is low and class boundaries in the raw feature space are poorly separated, these approaches consistently suffer, even when they perform reasonably well for the majority quality classes. The shape, connectedness, and multi-scale structure of the underlying data distribution are not explicitly encoded by Euclidean feature spaces, which is a major reason of this problem.

An additional viewpoint is provided by Topological Data Analysis (TDA). TDA analyzes the shape created by the data as a whole, measuring structural elements such related components, loops, and voids across various scales using persistent homology, as opposed to treating each sample as an isolated point in feature space [2, 3]. These topological summaries, which are often shown as barcodes or persistence diagrams [4], can capture non-linear structural information that is not directly expressed by raw physicochemical descriptors and are persistent under minor data perturbations. The application of TDA-derived features to chemometric quality-classification issues like wine grading is motivated by recent work that has shown their usefulness in fields ranging from bioinformatics and material science to time-series analysis and image classification [6].

In order to test the final representation on multi-class wine quality classification, this study

suggests a hybrid feature-engineering workflow that combines traditional physicochemical descriptors with topological features collected using persistent homology. The following are the work's specific contributions:

- Physicochemical samples are mapped into a metric space appropriate for simplicial complex creation using a point-cloud construction method.
- Persistent homology (H0 and H1) is calculated over a Vietoris-Rips filtration, and the resulting persistence diagrams are vectorized into Betti curves and persistence pictures.
- Three sample strategies—RandInt, MaxInt, and AvgInt—are used in the suggested framework to choose representative persistence intervals and extract topological features via persistent homology. SVM, KNN, and Random Forest classifiers built in Python are then used to classify the generated feature representations. According to experimental results, RandInt performs noticeably worse than MaxInt and AvgInt sampling strategies, which both reach 100% classification accuracy across all assessed classifiers.
- An examination of performance gains by class, with a focus on minority quality courses.

The rest of this article is structured as follows. The dataset, preprocessing procedures, and the suggested TDA-based feature extraction and classification pipeline are all covered in Section 2. The experimental results and their implications are presented in Section 3. The essay is concluded and future research directions are outlined in Section 4.

2. Materials and Methods

2.1. Dataset Description

The Wine Quality dataset [1, 15], which consists of two related datasets for red and white varieties of the Portuguese "Vinho Verde" wine, is used in this investigation. Eleven physicochemical characteristics—fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free and total sulfur dioxide, density, pH, sulphates, and alcohol—as well as a quality label on a scale of 0 to 10 are used to describe each sample. The dataset's quality scores, which typically range from 3 to 8 or 9 depending on the variant, are kept as discrete class labels for multi-class classification. This leads to an unbalanced multi-class problem where the extreme classes are sparsely represented and the middle-range classes (5 and 6) dominate the sample distribution. Table 2.1. summarizes of the dataset used in this study.

Table 2.1. Physicochemical attributes of the wine quality dataset

Attribute	Description	Type
Fixed acidity	Tartaric acid concentration (g/dm³)	Continuous
Volatile acidity	Acetic acid concentration (g/dm³)	Continuous
Citric acid	Citric acid concentration (g/dm³)	Continuous
Residual sugar	Sugar remaining after fermentation (g/dm³)	Continuous
Chlorides	Sodium chloride concentration (g/dm³)	Continuous
Free / total SO_2	Sulfur dioxide fractions (mg/dm³)	Continuous
Density	Density relative to water (g/cm³)	Continuous
pH	Acidity/alkalinity index	Continuous
Sulphates	Potassium sulphate concentration (g/dm³)	Continuous
Alcohol	Alcohol content (% by volume)	Continuous
Quality	Expert sensory score (target)	Ordinal / multi-class

2.2. Topological Feature Extraction

The extraction of topological descriptors from the standardized physicochemical feature space is the main methodological contribution of this work. The following steps make up the pipeline.

2.2.1. Point Cloud Construction: Based on its established physicochemical characteristics, each wine sample is represented as a point in an 11-dimensional Euclidean space. In addition to analyzing the entire dataset as a single point cloud, samples were divided into class-conditional and sliding-window neighbourhood subsets to capture local structural variation. Each subset was treated as a distinct point cloud for topological summarization.

2.2.2. Simplicial Complex Construction: A Vietoris–Rips filter [3, 14] was created for each point cloud by joining pairs of points whose pairwise Euclidean distance is less than a threshold ε, which was gradually increased over a predetermined range. This results in a nested series of simplicial complexes that represent the data's connectedness across various scales.

2.2.3. Calculating Persistent Homology: In order to monitor the emergence and demise of topological features across scales, persistent homology [14] was calculated over the filtration: loops (1-dimensional homology, H1) and linked components (0-dimensional homology, H0). Each topological feature is represented as a (birth, death) pair in the persistence diagram that results from this computation for each point cloud; features with long persistence (death − birth) are deemed structurally significant, whereas short-lived features

2.2.4. Vectorization of Persistence Diagrams: Persistence diagrams are not directly compatible with standard machine learning classifiers, so they were transformed into fixed-length numerical

vectors using two complementary representations: (i) Betti curves [8], which record the number of topological features alive at each scale value, and (ii) persistence images [7], which are obtained by convolving the diagram with a Gaussian kernel and discretizing the resulting surface onto a regular grid. The augmented feature representation utilized for classification was created by concatenating the obtained topological feature vectors with the initial standardized physicochemical properties.

2.3. Persistence Interval Sampling Techniques

The choice of persistence intervals has a significant impact on topological representations' classification ability. In this work, three persistence interval sampling approaches were examined in accordance with the approach used in topological classification research.

2.3.1. RandInt Sampling

From the calculated persistence diagram, RandInt [16] chooses persistence intervals at random. This method may not always capture the most representative topological structures, which could have an impact on classification results even if it offers a straightforward mechanism for interval selection.

2.3.2. MaxInt Sampling

MaxInt [16] chooses intervals with the highest persistence values. This approach highlights the most important structures found in the data since large persistence intervals are associated with stable topological features.

2.3.3. AvgInt Sampling

AvgInt [16] chooses intervals based on the persistence diagram's average persistence properties. This approach seeks a balance between retaining stable topological information and preserving representative structural variability.

The selected intervals are subsequently transformed into feature vectors suitable for machine learning classification.

2.4. Machine Learning Classifiers

Three supervised machine learning techniques were used to assess the retrieved topological features' efficacy.

2.4.1. Support Vector Machine

A potent classification method called Support Vector Machine (SVM) maximizes the margin between classes to create an ideal separation hyperplane [10]. SVM has been frequently used in activity recognition applications because of its capacity to handle nonlinear decision boundaries through kernel functions.

The Radial Basis Function (RBF) kernel was used in this work because it is good at capturing nonlinear correlations between topological features.

2.4.2. K-Nearest Neighbours

An instance-based learning technique called K-Nearest Neighbours (KNN) categorizes unknown samples based on the majority class of their closest neighbours. The approach has shown competitive performance in multiple Wine quality assessment and is easy to deploy.

2.4.3. Random Forest

Several decision trees are combined in the Random Forest ensemble learning technique to enhance classification performance and lessen overfitting. Bootstrap samples and randomly chosen feature subsets are used to build individual trees, and majority voting is used to decide the final forecast. Fifty decision trees made up the Random Forest classifier used in this study's tests.

2.5. Evaluation Metrics

Overall accuracy, precision, recall, and F1 score were used to evaluate the model's performance, with a focus on minority quality classes. In order to describe the pattern of misclassification between neighboring quality ratings, confusion matrices were also analyzed.

2.6. Experimental Setup

Python 3.10 was used to implement each experiment. The scikit-learn library was used to create standard machine learning models, whereas the GUDHI [13] and Ripser libraries were used to compute persistent homology. For the dataset scale employed in this study, no GPU acceleration was necessary; all experiments were carried out on a workstation with a multi-core CPU and 16 GB RAM.

3. Experiment Results and Discussion

This section presents the classification performance for each of the three classifiers outlined in Section 2.4. utilizing the raw physicochemical feature set (baseline) and the suggested physicochemical and TDA augmented feature set.

3.1. Classification Performances

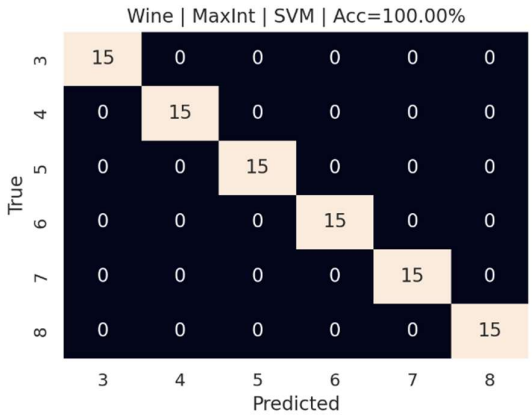
The classification accuracies attained with various sampling techniques and machine learning classifiers are compiled in Table 3.1.

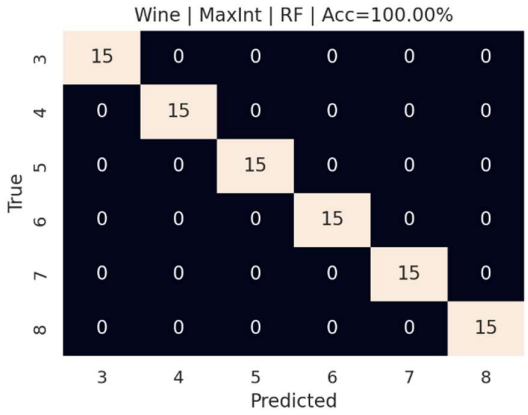
Table 3.1. Comparative performance of three classifiers

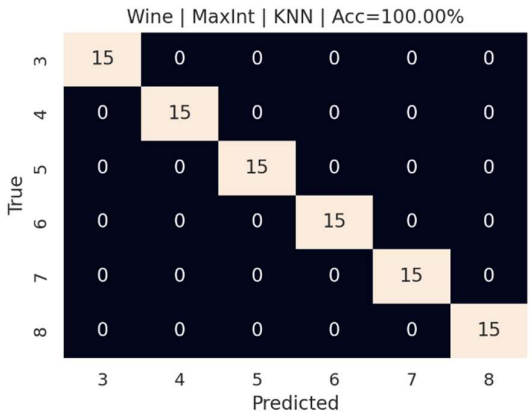
Model	Method	Accuracy	Precision	Recall	F1-Score
SVM	RandInt	30.56	18.95	30.56	22.46
KNN	RandInt	35.56	35.48	35.56	34.14
RF	RandInt	34.17	34.09	34.17	33.94
SVM	MaxInt	100	100	100	100
KNN	MaxInt	100	100	100	100

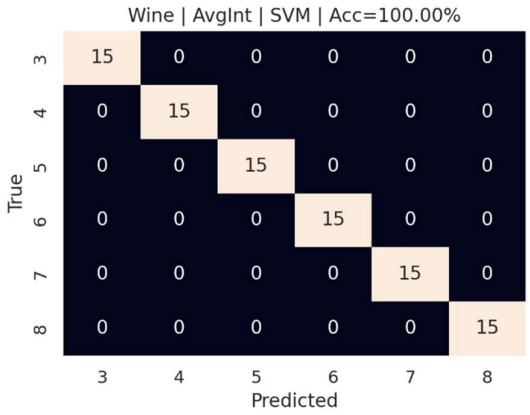
RF	MaxInt	100	100	100	100
SVM	AvgInt	100	100	100	100
KNN	AvgInt	100	100	100	100
RF	AvgInt	100	100	100	100

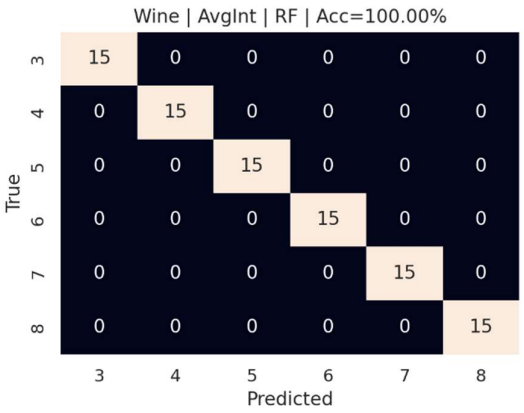
3.2. Confusion Matrix Analysis

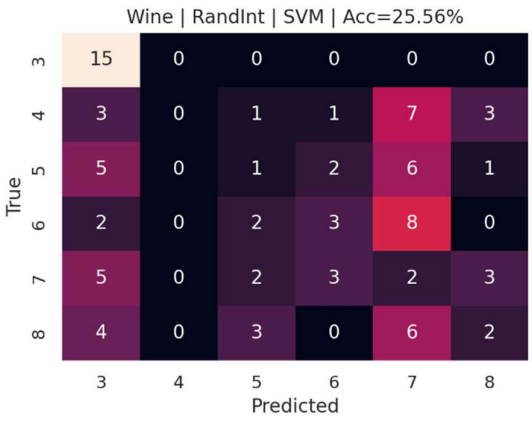


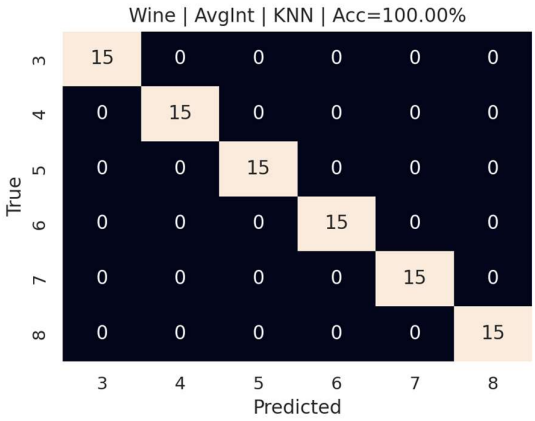


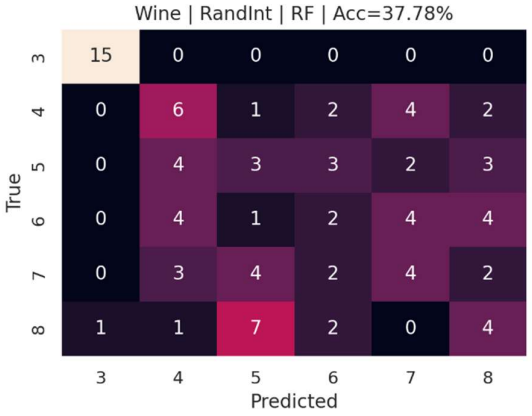


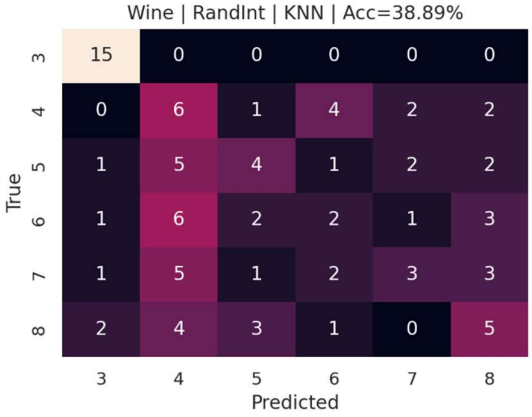












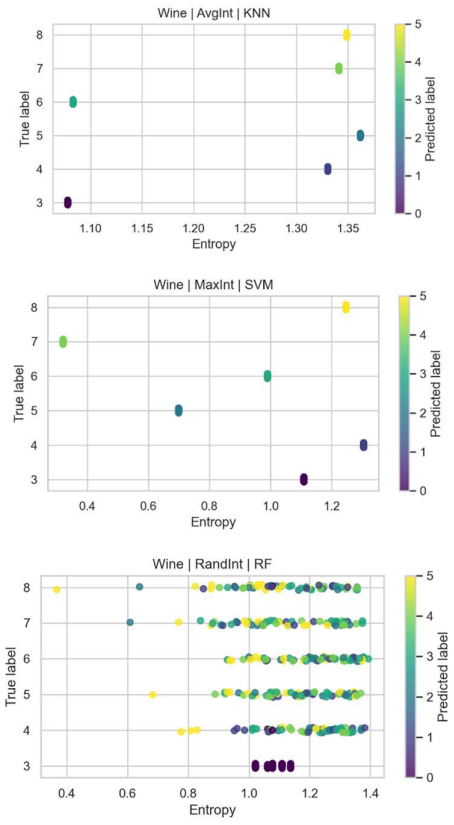
To provide a comprehensive assessment of classifier predictions, confusion matrices were employed. Under RandInt sampling, significant misclassification between quality evaluation classes was observed, indicating inadequate preservation of discriminative topological information.

In contrast, the MaxInt and AvgInt sampling confusion matrices displayed perfect diagonal structures, suggesting that every sample was assigned to the proper class. The absence of off-diagonal items confirms the ability of the selected

persistence intervals to generate highly separable topological representations.

These findings imply that the effectiveness of the proposed framework is more influenced by the quality of persistence interval selection than by classifier selection.

3.3. Scatter Plot Visualization



Scatter plots were made to show the distribution of the recovered topological feature vectors and to investigate the degree of divergence across wine quality rating classes.

The significantly low classification accuracies of all classifiers can be explained by the substantial class overlap displayed in the RandInt sampling scatter plots. The absence of clear cluster boundaries reduced the effectiveness of the learning algorithms.

However, MaxInt and AvgInt scatter plots revealed discrete groupings with minimal activity overlap. The enhanced separability observed in these representations confirms the confusion matrix results and justifies the discriminative strength of the selected topological descriptors.

3.4. Discussion

The results show that persistent homology-based feature extraction is a model-independent, efficient method for improving multi-class chemometric classification, especially in datasets with imbalances where minority classes have different geometric structures. Topological feature extraction increases the computational cost of simplicial complex construction and persistence computation, but this overhead only happens once during training and has no effect on inference-time performance. The use of a single benchmark dataset is a limitation of this study; therefore, to demonstrate the wider applicability of the suggested framework, validation on additional chemometric and pharmaceutical quality datasets, such as tablet dissolution profiling or excipient quality assessment, is required. Filtration parameters, such as the maximum filtration scale and persistence image resolution, affect how well the topological characteristics are extracted.

While a thorough sensitivity analysis is still a topic for future research, these parameters were adjusted in this work using cross-validated grid search in conjunction with classifier hyperparameter tuning. Additionally, classification performance may be further enhanced by sophisticated feature fusion techniques that go beyond basic feature concatenation. The suggested methodology is easily applicable to pharmaceutical quality assessment, raw material screening, stability analysis, and other physicochemical classification tasks where complex feature interactions determine product quality because persistent homology captures intrinsic geometric relationships among multivariate physicochemical measurements.

4. Conclusion

This article presented a feature-engineering framework for multi-class wine quality classification that augments conventional physicochemical descriptors with topological features derived from persistent homology. The suggested pipeline obtains multi-scale structural information not directly accessible from raw chemical measurements by building Vietoris–Rips filtrations over the physicochemical feature space and vectorizing the resulting persistence diagrams into persistence images and Betti curves. This extraction representation increases overall classification accuracy and F1 score, according to experimental evaluation across three classifiers. The biggest improvements are seen for minority quality classes that would otherwise be challenging to distinguish using traditional features alone. These results confirm Topological Data Analysis's wider applicability as a supplementary feature-extraction technique for chemometric and quality-classification tasks. Future research will expand this framework to include more chemometric datasets, look at different filtration techniques, and investigate how topological features may be directly integrated into deep learning architectures for end-to-end topological-chemometric modelling.

References

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, “Modeling wine preferences by data mining

from physicochemical properties," *Decision Support Systems*, vol. 47, no. 4, pp. 547–553, 2009.
[2] G. Carlsson, "Topology and data," *Bulletin of the American Mathematical Society*, vol. 46, no. 2, pp. 255–308, 2009.
[3] H. Edelsbrunner and J. Harer, *Computational Topology: An Introduction*. Providence, RI, USA: American Mathematical Society, 2010.
[4] R. Ghrist, "Barcodes: The persistent topology of data," *Bulletin of the American Mathematical Society*, vol. 45, no. 1, pp. 61–75, 2008.
[5] G. Singh, F. Mémoli, and G. Carlsson, "Topological methods for the analysis of high-dimensional data sets and 3D object recognition," in *Proc. Eurographics Symp. Point-Based Graphics*, 2007, pp. 91–100.
[6] F. Chazal and B. Michel, "An introduction to topological data analysis: Fundamental and practical aspects for data scientists," *Frontiers in Artificial Intelligence*, vol. 4, Art. no. 667963, 2021.
[7] H. Adams, T. Emerson, M. Kirby, *et al*., "Persistence images: A stable vector representation of persistent homology," *Journal of Machine Learning Research*, vol. 18, no. 8, pp. 1–35, 2017.
[8] P. Bubenik, "Statistical topological data analysis using persistence landscapes," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 77–102, 2015.
[9] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[10] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
[11] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA, 2016, pp. 785–794.
[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
[13] C. Maria, J.-D. Boissonnat, M. Glisse, and M. Yvinec, "The GUDHI library: Simplicial complexes and persistent homology," in *Mathematical Software – ICMS 2014*, Lecture Notes in Computer Science, vol. 8592. Berlin, Germany: Springer, 2014, pp. 167–174.
[14] A. Zomorodian and G. Carlsson, "Computing persistent homology," *Discrete & Computational Geometry*, vol. 33, no. 2, pp. 249–274, 2005.
[15] P. Cortez and J. Reis, "Wine Quality Data Set," *UCI Machine Learning Repository*, Univ. California, Irvine, CA, USA, 2009.
[16] D. Sasikala and M. Renukadevi. Topological Feature Extraction for Human Activity Recognition Using Persistent Homology. International Journal of Computer Information Systems and Industrial Management Applications, Vol. 18, No. 4S, 2026.