University of Barishal

Thesis Progress

Supervisor

Dr. Md Manjur Ahmed
Associate Professor,
Dept. of CSE,
University of Barishal

Submitted by:

Sarna Das (20CSE001)
Session: 2019-20
Year: 4th
Semester: 2nd
Submission Date: 16th September 2025

1. Problem Statement

A time series is an ordered sequence of data points collected over time. Examples include electrocardiograms (**ECG**) in medicine, stock prices in finance, temperature readings in climate science, and sensor readings in industrial processes. Time series analysis is crucial for detecting patterns, predicting trends, and inferring underlying processes. Among the many analysis tasks, motif discovery is arguably the most fundamental because motifs tend to manifest important repetitive events or behaviors. Despite its importance, motif discovery has a number of long-standing issues including variable motif lengths, noise and distortions, multidimensional complexity, parameter sensitivity, and computational scalability. Several approaches to address parts of this problem have been proposed over the years.

The initial breakthrough was made by Semantic Motifs (**Keogh et al., ICDM**), which made it possible to have don't-care regions between a prefix and a suffix. This enabled the discovery of higher-level patterns where the middle portion may differ, and it was found to be applicable in domains like animal behavior and speech analysis. However, the method required manual tuning of parameters (prefix/suffix lengths and don't-care region size) and was only applicable to univariate time series, which limited its broader acceptance.

Next, the Persistence-based approach (**PEPA/A-PEPA, Oudre et al., 2024**) solved the issue of arbitrary similarity thresholds. Using persistent homology, motifs were represented as connected subgraphs in similarity graphs of subsequences. This allowed for variable-length motif detection without pre-defined thresholds, while the adaptive version (A-PEPA) was even capable of automatically estimating the number of motifs. The method was noise-robust and motif-length flexible but computationally expensive and focused primarily on univariate datasets.

In contrast, the **LDSA** method (2024) was aimed at improving efficiency for motif mining. It converted time series to symbolic representations via Piecewise Aggregate Approximation (**PAA**) and Symbolic Aggregate approXimation (**SAX**) and constructed a suffix array with the **DC3** algorithm for efficient searching. With edit distance tolerance incorporated, **LDSA** could mine longer and approximate motifs that were missed by earlier approaches. Symbolic approximation, however, carried the danger of losing fine-grained temporal information, while parameter tuning was still needed.

More recently, the **ILMGD** framework (2024) pushed motif discovery into the realm of multidimensional time series. By leveraging matrix profiles with index-link merging, **ILMGD** discovered motif groups in parallel across many signals. This made it tractable to modern multivariate datasets, such as IoT sensor data or biomedical recordings. However, it remained vulnerable to distance measures and could not scale to very long or streaming datasets.

In summary, motif discovery research evolved incrementally like semantic motifs introduced flexibility, persistence-based motifs introduced robustness, suffix-array methods introduced efficiency, and matrix profiles introduced multidimensional scalability. There is no method that contains all these strengths, and an open problem still exists like the formulation of a unifying framework encompassing robustness, flexibility, efficiency, scalability, and multidimensional support.

Several methods have addressed parts of this problem, each with its strengths and weaknesses:

Approach / Paper	Strengths	Limitations
Semantic Motifs (Keogh et al., ICDM)	Captures **prefix-suffix motifs with don't-care regions**; useful in behavior and speech analysis; efficient $O(n^2)$ algorithm.	Parameters (prefix/suffix lengths, max don't-care length) must be defined; limited to univariate time series
Persistence-based Motif Discovery (PEPA/A-PEPA, Oudre et al., 2024)	Uses **topological persistence** to detect motifs that persist across thresholds; handles variable-length motifs; threshold-free	Computationally expensive for large datasets. Requires manual parameter tuning and limited to univariate data
Suffix Array Index Approach (LDSA, 2024)	Efficiently detects **variable-length motifs** using **symbolic representation (PAA + SAX)** and edit distance tolerance; supports longer patterns	May lose fine-grained details due to symbolic approximation; parameters still needed
ILMGD: Multidimensional Motif Group Discovery (2024)	Detects **motif groups** in multidimensional time series using matrix profiles; scalable across multiple signals	Sensitive to distance metric; performance may degrade on long sequences or streaming data

2. Research Gap

While each approach addresses some challenges (robustness, flexibility, efficiency, or multidimensionality), no existing method integrates all these strengths in a single framework. This gap motivates a comprehensive study to design a unified, scalable, and interpretable motif discovery system. From the literature, several gaps remain unaddressed:

1. **Unified Framework** – Existing methods address either robustness (persistence), flexibility (semantic motifs), efficiency (suffix array), or multidimensionality (matrix profile), but none combine all four.
2. **Automatic Parameter Tuning** – Most methods still rely on manual selection of motif length, thresholds, or edit distance, which limits real-world applicability.
3. **Multivariate and Streaming Time Series** – Few methods are capable of discovering motifs in high-dimensional or continuously streaming data.
4. **Interpretability** – Motif discovery is often a black-box process; domain experts (doctors, engineers) need interpretable results for actionable insights.
5. **Domain-Specific Adaptation** – Existing approaches are mostly general; tailored solutions for healthcare, climate, or industrial datasets are still limited.

3. Challenges

As a beginner, working on motif discovery research, I encountered:

1. **Understanding the Context of Paper:** As a beginner, it feels hard to understand the terminology, algorithms and other contexts of a paper**.**
2. **Complex Mathematics** – Persistent homology, edit distance optimizations, and matrix profiles require a strong understanding of mathematical foundations.
3. **Implementation Complexity** – Efficiently coding these algorithms for large datasets is non-trivial and requires optimization expertise.
4. **Data Issues** – Large data handling feels tough.
5. **Result Interpretability** – Making motifs understandable and meaningful to end-users is an additional layer of difficulty.

4. Objectives

The research aims to:

1. **Develop a Motif Discovery Framework** – A framework that will try to overcome some major limitations of previous motif discovery algorithms.
2. **Automate Parameter Selection** – Use data-driven or learning-based techniques to determine motif length, thresholds, and edit distance dynamically.
3. **Evaluate Scalability and Accuracy** – Testing on large, real-world and streaming datasets for runtime and motif quality.
4. **Apply to Real-World Domains** – Validation of effectiveness on healthcare signals, climate monitoring and industrial sensor data.
5. **Enhance Interpretability** – To provide visualizations and domain explanations to make motifs actionable for practitioners.

5. Literature Review

The first of the four techniques is the work of Keogh and colleagues (**ICDM**) on semantic motifs. This technique introduced prefix-suffix motifs with arbitrary don't-care regions in the middle to enable the discovery of higher-level structure in time series at multiple levels beyond simple exact subsequence matching. Semantic motifs have proven to be helpful in applications such as animal behavior analysis and speech recognition, where recurring patterns do not necessarily match exactly. The method employs an efficient **$O(n^2)$** motif discovery algorithm, although its major drawback is the requirement for manual specification of prefix length, suffix length, and acceptable don't-care region, potentially hindering usability in large-scale or automated applications.

Building on motif discovery challenges, the Persistence-based approach (**PEPA/A-PEPA**, Oudre et al., 2024) leveraged persistent homology to encode motifs as connected subgraphs of subsequence similarity graphs. Unlike fixed-threshold approaches, **PEPA** is capable of discovering variable-length motifs without predefined similarity thresholds. Its adaptive version,

A-PEPA, goes one step further and automatically determines the number of motifs through persistence diagrams. This makes the method robust to noise and motif length detection flexible. However, it remains computationally expensive and largely restricted to univariate time series.

Around the same time, writers introduced the **LDSA** method (2024) for variable-length motif mining through symbolic indexing techniques. The technique transforms the raw time series to symbolic representation using Piecewise Aggregate Approximation (**PAA**) and Symbolic Aggregate **approXimation (SAX)** and constructs a suffix array via the DC3 algorithm for efficient searching. To enhance robustness, it incorporates edit distance, which enables slight variations in motifs and enables the discovery of longer approximate patterns. While efficient, the symbolic representation risks losing fine-grained temporal features, and its parameters (e.g., edit distance thresholds) need to be carefully tuned.

Very recently, the **ILMGD** framework (2024) has extended motif discovery to the setting of multidimensional time series. By employing matrix profiles under index-link merging, **ILMGD** discovers groups of motifs under high-dimensional configurations, making it amenable to sensor-rich and multivariate data. The algorithm is scalable with improved efficiency compared to most earlier algorithms, and it takes the application of motifs from univariate analysis further as well. Nevertheless, its effectiveness depends upon the choice of the distance measure, and scaling to extremely large or streaming data remains a challenge.

Together, these approaches signify the path of motif discovery research: from semantic flexibility (Keogh) to robustness through topology **(PEPA/A-PEPA),** efficiency through symbolic suffix arrays (**LDSA**), and finally multidimensional scalability (**ILMGD**). Nevertheless, no single approach fully meets the combined demands of robustness, flexibility, efficiency, multidimensionality, and parameter-free discovery, thereby leaving an important gap for further research exploration.

6. Methodology

Step 1: Dataset Collection & Preprocessing

- To gather benchmark datasets: ECG/EEG, climate, industrial sensors, finance.
- To normalize, handle missing data, and prepare univariate and multivariate series.

Step 2: Framework Design

- Combine the features of **topological persistence** (robust motif detection), **semantic motif structure** (prefix-suffix flexibility), **suffix-array indexing** (efficiency), **matrix profile** for multidimensional support.

Step 3: Algorithm Implementation

- Symbolic preprocessing: PAA + SAX.
- TO construct suffix arrays with DC3; integrate edit distance for motif tolerance.
- To use persistence diagrams to detect robust motifs and semantic patterns.

- Automatic parameter estimation via clustering or heuristics.

Step 4: Evaluation

- Compare with PEPA/A-PEPA, Semantic Motif Finder, ILMGD, LDSA.
- Metrics: motif overlap, DTW distance, motif length, runtime, scalability.

Step 5: Application

- Healthcare: ECG/EEG anomaly detection.
- Climate: Identify recurring weather or environmental patterns.
- Share-market: To identify changes.

Step 6: Visualization & Interpretability

- Interactive dashboards for motif visualization.
- Explain motifs in domain-specific contexts for actionable insights.

7.References

1. Oudre, L., Truong, C., & Germain, M. (2024). *Persistence-based Motif Discovery in Time Series*. IEEE Transactions on Knowledge and Data Engineering (TKDE).
2. Imani, S., & Keogh, E. (2018). *Time Series Semantic Motifs: A New Primitive for Finding Higher-Level Structure in Time Series*. ICDM.
3. Zhang, et al. (2024). *Multidimensional Time Series Motif Group Discovery Based on Matrix Profile*. Knowledge-Based Systems.
4. Wang, J., Zhu, Y., & Sun, J. (2024). *Variable Length Motif Mining Method Based on Suffix Array Index*. AIHCIR 2024.