

Adaptive Leverage-Score Sketching and Sparse Random Projections for Scalable Streaming Clustering with Concept x Drift

Wasnaa Hadi Ghasab¹ , Saif.A.H Moamin² , Alyaa Hussein Ali³ , Rusul Mansoor Al-Amri^{2*} ,
Liaw Geok Pheng⁴ 

¹ College of Management and Economics, University of Kerbala, Kerbala, Iraq.

² Department of information technology, Collage of computer science and information technology, University of Kerbala.

³ Collage of Dentistry, University of Kerbala. Kerbala, Iraq.

⁴ Faculty of Teknologi Maklumat dan Komunikasi (FTMK), Universiti Teknikal Malaysia Melaka (UTeM), 76100, Durian Tunggal, Melaka, Malaysia.

*Corresponding Author: Rusul Mansoor Al-Amri

DOI: <https://doi.org/10.31185/wjcms.529>

Received 30 May 2026; Accepted 27 August 2026; Available online 30 September 2026

ABSTRACT: Clustering in high-dimensional streaming data is formidable due to concept drift, changing data distributions, and computational limitations. This work proposes an adaptive sketching method aimed to clustering these highly-dimensional data streams, which can simultaneously handle both scalability and concept drift. The framework combines leverage-score sampling, sparse random projections, adaptive rank selection, and online micro-clustering in a common incremental architecture. Updating the sketch representation benefits from an online learning process based on a reconstruction-error monitoring mechanism that dynamically adjusts the sketch representation in parallel to the changes in the underlying data distribution without the need for recomputing the entire sketch. Experimental results on three benchmark datasets (CoverType, ELEC, and Twitter) show that the proposed framework provides superior performance compared to state-of-the-art streaming clustering methods. It achieved ARI/NMI scores of 0.68/0.61 on CoverType, 0.72/0.65 on ELEC, and 0.39/0.42 on Twitter, surpassing CluStream, DenStream, StreamKM++, and LEMON. Moreover, the ablation study verifies the benefit of the proposed design, where the framework with full design achieves 0.68 in ARI, and that without adaptive rank selection shows 0.54 in ARI, suggesting that adaptive rank selection is a key for clustering performance. The detailed experiments demonstrate that the elaborated framework achieves optimal balance among clustering performance, adaptability and efficient computational cost, making it suitable for real-world applications such as IoT analytics, financial monitoring and online anomaly detection.

Keywords: Streaming Data Mining, Concept Drift, Adaptive Sketching, Streaming Machine Learning, Sparse Random Projections



1. INTRODUCTION

Data streams are becoming a primary source for certain real-time analytics due to a need for the continuous generation of data from Internet of Things (IoT) infrastructures, sensor networks, financial systems, social media platforms, and other digital services. Streaming data are different than traditional batch datasets; the data arrive in a continuous fashion and may be too large to store or compare to repeatedly process completely. Hence, stream-mining algorithms operate with a limited memory and computing resource budget in the settings of incremental model updates. Stream clustering is especially difficult because meaningful groups should be detected within limited time and their representations are gradually changed as the data evolve over time. Compact summarization of micro-clusters for scalable clustering over continuously arriving data, as demonstrated in early stream-clustering frameworks, reveals the challenge in maintaining quality clustering with non-stationary distributions over time [1-

2]. This problem becomes more complex for high-dimensional streaming data, in which the distinct features exist as redundancy, correlation, and distinctiveness. High Dimensionality — In high-dimensional spaces, the cost of evaluating distance, making a covariance guess, matrix decomposition, clustering, and keeping the data compact with some mean of organism the most significant portion of the space increases. As such, dimensionality reduction and low-rank approximation represent two key pieces of the puzzle, by allowing us to reduce the computational footprint while preserving important information. Research in randomized numerical linear algebra has led to efficient randomized algorithms for large matrices using techniques including random projections, sampling and randomized low-rank matrix approximation, all of which allow for the processing of large matrices far more efficiently than with traditional deterministic decompositions. [2]. These approaches are particularly attractive for streaming applications because they can construct compact representations without repeatedly processing the complete historical dataset [2]. Matrix sketching provides a natural framework for maintaining compact representations in data-stream environments [3]. Frequent Directions, for example, processes a matrix incrementally and maintains a bounded-size sketch with provable approximation guarantees for low-rank structure [3]. Sparse random projections provide a complementary mechanism for reducing the computational cost of high-dimensional transformations while preserving distance-related information [4]. Incremental data-stream sketches based on sparse random projections have also demonstrated their potential for online clustering of high-dimensional streams [5]. However, efficient representation alone does not resolve the challenge of continuously changing data distributions[11].

In real-world streaming environments, the statistical characteristics of incoming data may change over time. Such changes are commonly referred to as concept drift or, in unsupervised settings, evolving data distributions [5],[6]. Drift may occur abruptly or gradually and may involve recurring or structural changes in the data-generating process. Consequently, a representation or model that is effective under one distribution may become less appropriate as the stream evolves. Concept-drift research has therefore emphasized detecting changes and adapting learning systems to new distributions [5],[6]. Recent literature has also emphasized the ongoing challenges of recurring and non-stationary changes in stream learning [7]. Specific challenges of clustering come from the fact that, unlike classification, adaptation has generally to rely on continuously available class labels. At the same time, changes are more likely to appear in the forms of changes in cluster centroids, covariance structures, densities, or cluster proportions, and the relationships among features. Existing algorithms for stream-clustering deal with the problem of evolving data using compact summaries and incremental updates [10]. CluStream proposed an online micro-cluster based framework for finding evolving streams [1], and later on, DenStream established a density-based micro-cluster representation of evolving clusters and identifying potential clusters from outliers [9]. While these methods lay important groundwork for scalable streaming clustering, it becomes more difficult to -maintain an informative representation when data are high-dimensional and nonstationary.

This raises an important question not just on how to cluster a stream as it evolves but also on maintaining an up to date representation as the stream evolves. While fixed-dimensional projection or fixed-rank sketches are computationally efficient, their representational capacity may become incompatible with the changing underlying structure. Cancelling potentially meaningful structural information when it is small, or increasing computational and memory costs when it is needlessly large. This provides a quality scalability trade-off. Thus; an adaptive streaming representation should adjust its effective rank according to the information content of the current stream. The second challenge revolves around which observations we use in order to build the sketch. In many scenarios, observations may not have comparable importance; uniform sampling assumes they do. Leverage scores are a data-dependent notion that measure influences of rows on spectral structure of the matrix and have been employed in randomized matrix approximation and sampling [8]. For the case of continuously arriving multidimensional data, online leverage score sampling has been studied as a mechanism for selecting a few informative observations from an entire stream of data without needing to process the whole stream [8]. But the covariance and distributional structure is evolving, so the importance of observation also changes. Therefore, we will require that, for non-stationary streams, this sampling priorities from the leverage-score mechanism be updated in conjunction with shifts in the underlying data. These challenges reveal a broader methodological issue. Dimensionality reduction, sketch construction, drift adaptation, and clustering are often treated as related but partially separate components. Under high-dimensional concept drift, however, changes in the data distribution may simultaneously affect which observations are informative, which directions dominate the data, how many dimensions are required to represent the current structure, and how clusters should be updated. Treating these processes independently can therefore create a mismatch between the representation supplied to the clustering algorithm and the current structure of the stream. The challenge is consequently not simply to detect drift after it occurs, but to allow the data representation itself to participate in the adaptation process. Despite substantial progress in stream clustering, randomized matrix approximation, leverage-score sampling, sparse projections, and concept-drift adaptation, an important gap remains in their joint treatment for scalable high-dimensional stream clustering. Existing research provides strong solutions for individual aspects of the problem: stream-clustering algorithms maintain compact cluster summaries [1],[10], randomized matrix methods provide efficient low-rank representations [2],[4], sparse projections reduce the

computational burden of high-dimensional transformations [4], and concept-drift research provides mechanisms for adapting learning systems to evolving distributions [6],[8]. However, the integration of data-dependent sampling, sparse dimensionality reduction, adaptive representation rank, and online clustering within a single incremental adaptation loop remains insufficiently explored. This gap is particularly important when the stream is simultaneously high-dimensional and non-stationary. To address this gap, this study proposes an Adaptive Leverage-Score Sketching and Sparse Random Projections framework for Scalable Streaming Clustering with Concept Drift. The framework integrates four complementary mechanisms within a unified incremental architecture: (i) adaptive ridge leverage-score sampling to prioritize observations influential to the evolving spectral structure; (ii) sparse random projections to obtain computationally efficient low-dimensional representations; (iii) adaptive rank selection based on reconstruction-error monitoring and an energy criterion; and (iv) online micro-clustering using an adaptive Mahalanobis metric to account for evolving cluster structure. Rather than reconstructing the complete representation whenever the stream changes, the proposed framework updates covariance information, sampling probabilities, the sketch, and its low-rank representation incrementally. Reconstruction error provides a feedback signal for adjusting the representation rank, thereby coupling representation adaptation with the evolving characteristics of the stream. Thus, the main contributory aim of this study is to design and validate the efficacious performance of a scalable streaming-clustering framework which preserves informative low-dimensional representations even under the conditions of high dimensionality and simulation of evolving data distributions, all while keeping computational and memory costs low. In particular, it explores: whether adaptive leverage-score sampling with sparse random projections creates informative low dimensional sketches and can computationally efficiently; whether rank adaptation based on reconstruction error preserves a suitable low-rank representation as the underlying data distribution shifts; and whether adaptive sketching blended with online micro-clustering can outperform legacy streaming-clustering methods in both the quality and scalability of the resulting clusters.

The main contributions of this study are as follows:

1. Unified adaptive streaming architecture: A unified framework integrating adaptive leverage-score sampling, sparse random projections, adaptive rank selection, and online micro-clustering into a single incremental pipeline for high-dimensional streaming data.
2. Adaptive data-dependent sketch construction: An adaptive ridge leverage-score mechanism that prioritizes structurally influential observations, combined with sparse random projections for computationally efficient dimensionality reduction.
3. Dynamic representation adaptation: A reconstruction-error-based rank-selection mechanism that dynamically regulates sketch complexity in response to changes in the underlying data structure without repeated full reconstruction.
4. Comprehensive empirical evaluation: An evaluation on real-world streaming datasets, including CoverType, ELEC, and Twitter, against established streaming-clustering baselines using ARI and NMI, computational efficiency, scalability, and ablation analysis.

the proposed framework bridges scalable randomized matrix approximation and adaptive stream clustering by making the compact representation an active component of the adaptation process. By jointly adapting observation selection, projection, representation rank, and clustering statistics, the framework is designed for streaming environments in which both data volume and statistical structure may change continuously, with potential applications in IoT analytics, financial monitoring, online anomaly detection, and other real-time data analysis systems.

2. RELATED WORK

Existing works have concentrated on one or more of the clustering problems typically raised by streaming high-dimensional data with concept drift [10],[12]. This section is divided into three main categories: (1) randomized dimensionality reduction methods focusing on streaming; (2) algorithm adaptations for clustering considering concept drift; and (3) scalable low-rank approximation methods.

2.1 Randomized Dimensionality Reduction in Streaming Data

Dimensionality reduction is not a new subject area; however, randomized approaches have garnered interest in the machine learning community due to their ease-of-use and theoretical assurances[13]. This is primarily because of the Johnson-Lindenstrauss lemma which mathematically ensures that pairwise distances are preserved with high probability even after very high dimensionality reduction. These methods have been adapted further such that they achieve a higher sparsity and lower computational complexity potentially compromising accuracy. Incremental sketching algorithms like Frequent Directions [4] achieve provable error guarantees on low-rank approximations in the streaming context. Such approaches, however, tend to consider all the data points equally regardless of their relative significance. Leverage-score sampling resolves this weakness, as it focuses on the rows that have a large impact on the spectral characteristics of the data. They are mostly taken to advantage of the concept drift and recent work has applied leverage-score sampling to

streaming environments [8], although these methods usually rely on assumptions about the underlying distribution, which are not necessarily true in practice.

2.2 Concept Drift Detection in Clustering.

The concept drift detection and adaptation issue in the supervised learning context has been widely investigated [5], however, the unsupervised context presents special issues [12]. While CluStream [1] proposed the use of micro-clusters to provide incremental summaries of streaming data, it lacks mechanisms for distributional change detection in high dimensions [11]. Subsequent extensions, such as DenStream [9], introduced density-based clustering. However, these remain highly sensitive to dimensionality. Recent research has investigated hybrid methods of drift detection with dimensionality reduction. Indicatively, a high-speed streams two-dimensional representation was proposed in [10] thus ensuring effective drift monitoring. Nevertheless, these techniques usually separate drift detection and the clustering solution, which results in non-optimal solutions. Our framework, on the other hand, is only built to incorporate drift detection into the sketching process and rank adjustments are made to it.

2.3 Approximation Techniques that are Scalable and of Low-Rank.

One of the key instruments of processing high dimensional data is low-rank approximation. Alternative scalable algorithm, randomized algorithms [2], have been introduced to substitute traditional SVD, especially when working with large matrices. These approaches tend to take random projections or subsampling to develop imprecise bases of the matrix range of the input. The latest developments have seen the introduction of adaptive rank selection mechanisms [14]. As an example, [15] suggested a dynamic range finder that changes the approximation rank according to the residual errors. Nevertheless, this type of methods is mostly aimed at batch processing and cannot be applied to streaming. Other literature has concentrated on robustness to outliers [18] but fails to take up distributional changes over time.

As compared to the current solutions, our design integrates data-conscious sketching, sparse projections, and adaptive rank selection into an overall streaming solution. However, in comparison with [19], where we focus on a particular type of data, our approach is data-distribution-independent. In addition to this, in [3] the problem of nonconvex tensors is discussed, whereas we are interested in fast matrix approximations of clustering. The most important increased complexity is our integrated monitoring-and-adaptation loop, making it efficient at changing with concept drift in real time without compromising scalability.

3. BACKGROUND AND PRELIMINARIES

To establish the theoretical foundation for our adaptive sketching framework, we first review key concepts in low-rank approximation and randomized dimensionality reduction. These techniques form the building blocks for handling high-dimensional streaming data while maintaining computational efficiency.

3.1 Low-Rank Matrix Approximation

Given a data matrix $A \in \mathbb{R}^{n \times d}$ with n observations and d features, low-rank approximation seeks to find a matrix $\hat{A}_k$ of rank $k \ll \min(n, d)$ that minimizes the reconstruction error :

$$\|A - \hat{A}_k\|_F \quad (1)$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The optimal solution is given by the truncated Singular Value Decomposition (SVD) [4], where $\hat{A}_k = U_k \Sigma_k V_k^T$, with U_k and V_k containing the top k singular vectors and Σ_k the corresponding singular values. However, computing the exact SVD is prohibitively expensive for large-scale streaming data, necessitating randomized approximations [2],[13]. Randomized algorithms construct a sketch $Y = A\Omega$, where $\Omega \in \mathbb{R}^{d \times k}$ is a random projection matrix. The columns of Y span an approximate basis for the range of A , from which the low-rank factors can be efficiently computed. This approach reduces the computational complexity from $O(nd^2)$ to $O(ndk)$, making it suitable for streaming settings.

3.2 Leverage-Score Sampling

Rows do not contribute equally to the spectral properties of a matrix. One of the measures of the influence each row has is leverage scores, which measure its alignment with the strong singular vectors. Given a matrix A with SVD $A = U\Sigma V^T$, the leverage score of the i -th row of A is given by:

$$\ell_i = \|U(i, :)\|_2^2 \quad (2)$$

The rows which get a higher leverage score are more important in maintenance of the matrix structure. These scores are much more precise when it comes to approximations through sampling rows than uniform sampling [8]. Streaming to approximate leverage scores can be incrementally preserved [8], and allows sketching of data without needing to pass the full matrix.

3.3 Sparse Random Projections

Johnson-Lindstrauss (JL) lemma [4] is a lemma that ensures that pairwise distance can be maintained by random projections. Sparse variants (also known as sparse forms) lower computational costs by setting most entries of projection matrices $\Omega_{\text{sp}} = 0$ [4]. A widely used choice is the Achlioptas matrix, where.

$$\Omega_{ij} = \begin{cases} +1/\sqrt{s} & \text{with probability } s/2 \\ 0 & \text{with probability } 1-s \\ -1/\sqrt{s} & \text{with probability } s/2 \end{cases} \quad (3)$$

Here, s controls the sparsity level. Despite the reduced density, these projections retain theoretical guarantees while significantly accelerating computation.

3.4 Concept Drift in Streaming Data

This phenomenon is known as concept drift. Concept drift in the context of clustering can be described as the change in the distribution of the data over time. This may be reflected in the form of changes in the centroids of the clusters, changes in the variances of the clusters, or changes in the structure of the clusters. These changes can be detected by tracking the statistics such as reconstruction errors and cluster cohesion [5],[6]. The following section describes the interactions between the components of the system in the context of achieving the objectives of scalable clustering.

4. ADAPTIVE RANDOMIZED LOW-RANK APPROXIMATION FOR STREAMING DATA WITH CONCEPT DRIFT

The three components of the proposed system work in tandem to solve the problem of clustering high-dimensional streaming data. These include dynamic leverage score sampling, sparse random projection, and adaptive rank selection. These are concerned with finding a good way of representing the streaming data in lower dimensional space. Figure 1 illustrates the architecture of this system.

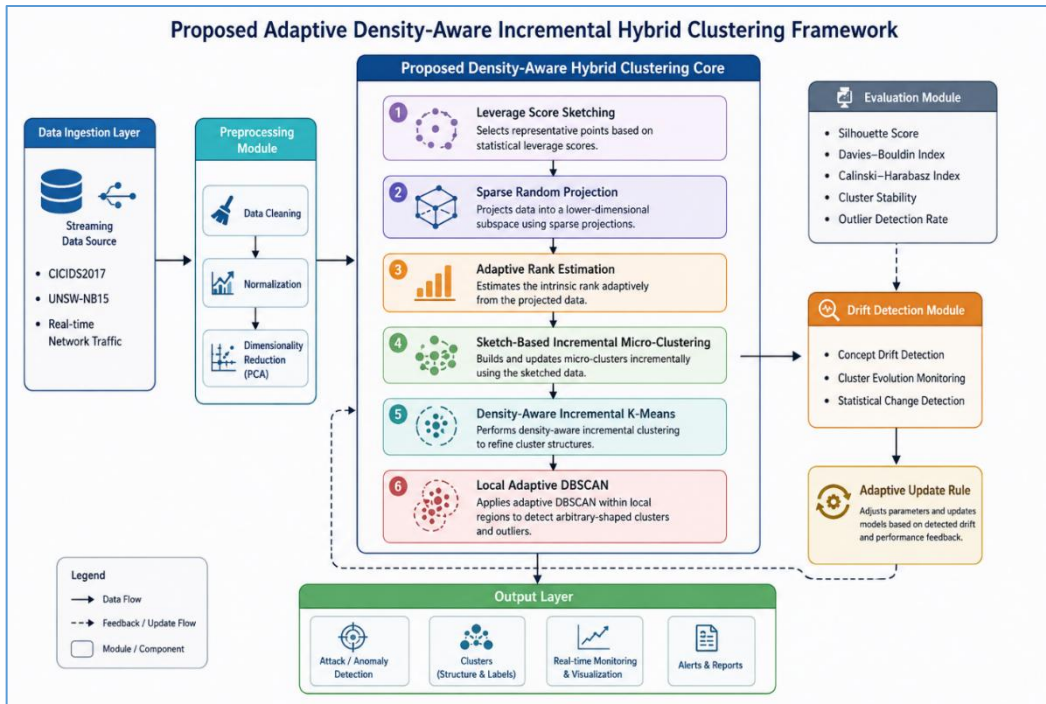


FIGURE 1. High-Level Workflow of Streaming Clustering with Adaptive Sketching

Figure-1 Instead, in this paper, we show the proposed adaptive sketching framework architecture for streaming clustering in Figure-1. A novel adaptive score sampling is used at the very start of the workflow to sample non-linear informative observations and dense random projection, sparse random projection is performed to efficiently reduce the dimensionality of datasets. The algorithm subsequently selects the inputs of the future by employing rank-selection strategies both for adaptation and online micro-clustering, whilst revising the monitoring-and-adaptation loop continually as a response to concept drift. Moreover, the integrated design can speed-up and improve the clustering quality by processing data variations on-the-fly.

4.1 Adaptive Leverage-Score Sampling for Concept Drift in Streaming Data

It proposes dynamic allocation of sampling probabilities to distributional changes to the data stream. Let, assume, $\mathbf{X}_t$, where t is a window of data. $\mathbf{X}_t \in \mathbb{R}^{n \times d}$ at time t , ridge leverage scores are computed to prioritize influential rows and mitigate multicollinearity. The i -th row score can be interpreted to mean:

$$\ell_i^{(t)} = \mathbf{x}_i^T (\mathbf{X}_t^T \mathbf{X}_t + \lambda \mathbf{I})^{-1} \mathbf{x}_i, \quad (4)$$

where λ is a regularization parameter that stabilizes the inverse computation. The sampling probability $p_i^{(t)}$ for each row is then proportional to its leverage score:

$$p_i^{(t)} = \frac{\rho_i^{(t)}}{\sum_{j=1}^n \rho_j^{(t)}}. \quad (5)$$

This guarantees that the rows that add more to the covariance structure are included in a better way in the sketch. Dealing with streaming updates is a non-trivial challenge and has led us to maintain an online estimate of the covariance matrix $\mathbf{C}_t = \mathbf{X}_t^T \mathbf{X}_t$ using exponential smoothing:

$$\mathbf{C}_t = \alpha \mathbf{C}_{t-1} + (1 - \alpha) \mathbf{x}_t \mathbf{x}_t^T, \quad (6)$$

where $\alpha \in (0,1)$ controls the decay rate. The inverse $(\mathbf{C}_t + \lambda \mathbf{I})^{-1}$ is updated incrementally via the Sherman-Morrison formula, avoiding costly recomputation.

4.2 Sparse Random Projections for Incremental Sketch Updates

To do so, we use sparse random projections to reduce the computational cost of dimensionality reduction while preserving pairwise distances, which are crucial for determining robust clusters in streaming settings[4]. The projection matrix $\mathbf{\Pi} \in \mathbb{R}^{d \times m}$ is constructed with sparsity parameter s , where each entry follows:

$$\Pi_{ij} = \begin{cases} +1/\sqrt{s} & \text{with probability } s/(2d) \\ 0 & \text{with probability } 1 - s/d \\ -1/\sqrt{s} & \text{with probability } s/(2d) \end{cases} \quad (7)$$

Here, $m \ll d$ is the target dimension, and s controls the trade-off between sparsity and projection quality. The sparsity ensures that each projection step requires only $O(s \cdot \text{nnz}(\mathbf{x}_t))$ operations for a new data point $\mathbf{x}_t$, where $\text{nnz}(\cdot)$ denotes the number of non-zero entries.

The projected sketch $\mathbf{S}_t \in \mathbb{R}^{n \times m}$ is updated incrementally as new data arrives. For each incoming point $\mathbf{x}_t$, its low-dimensional representation $\mathbf{y}_t = \mathbf{x}_t^T \mathbf{\Pi}$ is computed and appended to $\mathbf{S}_t$. To maintain a fixed sketch size, older rows are discarded or aggregated into micro-clusters. The sparse structure of $\mathbf{\Pi}$ ensures that this update is efficient even for high-dimensional data.

4.3 Adaptive Rank Selection through SVD Residual Monitoring

The proposed method dynamically adjusts the sketch rank k by monitoring the Frobenius norm error of the low-rank approximation[14]. For the current sketch $\mathbf{S}_t$, we compute the rank- k SVD $\mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^T$ and track the residual:

$$\epsilon_k^{(t)} = \|\mathbf{S}_t - \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^T\|_F \quad (8)$$

A smoothed threshold τ_t is maintained to detect significant deviations:

$$\tau_t = \beta \tau_{t-1} + (1 - \beta) \epsilon_k^{(t)} \quad (9)$$

where $\beta \in (0,1)$ controls the adaptation rate. When $\epsilon_k^{(t)}$ exceeds $\gamma \tau_t$ for a tolerance parameter $\gamma > 1$, the rank is adjusted incrementally. The new rank k' is selected as the smallest value satisfying:

$$\frac{\sigma_{k'+1}^2}{\sum_{i=1}^d \sigma_i^2} < \delta \quad (10)$$

where σ_i are the singular values of $\mathbf{S}_t$ and δ is a predefined energy threshold. This ensures the sketch retains sufficient variance while avoiding unnecessary dimensionality. The SVD is updated via a rank-one modification [15], avoiding full recompilation.

4.4 Sketch-Based Online Clustering with Adaptive Metrics

The low-dimensional sketch $\mathbf{S}_t$ is the basis for efficient online clustering. A convoluted way of saying, "We apply a Mahalanobis distance measure that is localized to the structure which the sketch represents." For a micro-cluster $\mathcal{C}_i$, we maintain the centroid $\mathbf{c}_i$ and covariance $\mathbf{M}_i$, which can be updated in an online fashion. An Important point is the distance between a new point $\mathbf{y}_t$ and $\mathcal{C}_i$ is computed as follows:

$$d(\mathbf{y}_t, \mathcal{C}_i) = \sqrt{(\mathbf{y}_t - \mathbf{c}_i)^T \mathbf{M}_i^{-1} (\mathbf{y}_t - \mathbf{c}_i)} \quad (11)$$

where $\mathbf{M}_i$ is estimated from the sketch data assigned to $\mathcal{C}_i$. The inverse covariance matrix is stabilized using ridge regularization, as defined in Equation (12):

$$\mathbf{M}_i^{-1} = (\mathbf{M}_i + \epsilon \mathbf{I})^{-1} \quad (12)$$

with ϵ ensuring numerical stability. This adaptive metric accounts for the evolving shape and orientation of clusters, unlike Euclidean distance which assumes spherical distributions.

When $\mathbf{y}_t$ arrives, it is assigned to the nearest micro-cluster if $d(\mathbf{y}_t, \mathcal{C}_i)$ falls below a threshold θ . Otherwise, a new micro-cluster is initialized. The centroids and covariances are updated via exponential smoothing:

$$\mathbf{c}_i \leftarrow \eta \mathbf{c}_i + (1 - \eta) \mathbf{y}_t \quad (13)$$

$$\mathbf{M}_i \leftarrow \eta \mathbf{M}_i + (1 - \eta) (\mathbf{y}_t - \mathbf{c}_i) (\mathbf{y}_t - \mathbf{c}_i)^T \quad (14)$$

where $\eta \in (0,1)$ controls the adaptation rate. This online update guarantees that the clustering model continues to react to concept drift but is relatively computationally efficient.

4.5 Unified Framework for Streaming and Drift Adaptation

The opportunity is essentially a reshuffling of the elements outlined in Sections 4.1-4.4 as a pipeline that takes in a new data stream, refreshes the model to remain current with a concept drift, and adjusts the various components to offset the shift in distribution. The innovation really comes through the extreme clarity with which tight coupling between sketching, rank adaptation and clustering with updates without full recompute are defined.

The following procedure is applied to data, updating the system for each incoming data point generated at every time step :

1. **Dynamic Leverage-Score Update:** The ridge leverage scores are recomputed using the current covariance estimate C_t from Equation 6. The sampling probability $p_i^{(t)}$ for each row in the sliding window is updated according to Equation 5.
2. **Sparse Projection:** The point x_t is projected into the low-dimensional space via $y_t = x_t^T \Pi$, where Π follows the sparse construction in Equation 7. The projection requires only $O(s \cdot \text{nnz}(x_t))$ operations due to sparsity.
3. **Sketch Maintenance:** The projected point y_t is appended to the sketch S_t , and the oldest point is removed to maintain a fixed window size. The sketch's SVD is incrementally updated using the method from [15].
4. **Rank Adaptation:** The residual error $\epsilon_k^{(t)}$ (Equation 8) is computed and compared against the smoothed threshold τ_t (Equation 9). If $\epsilon_k^{(t)} > \gamma \tau_t$, the rank k is adjusted according to the energy criterion in Equation 10.
5. **Online Clustering:** The point y_t is assigned to the nearest micro-cluster using the adaptive Mahalanobis distance (Equation 11). The centroids and covariance matrices are updated via Equations 13-14.

The sparse projection and incremental SVD steps are the computational components that dominate the computational complexity of every update and they are in $O(s \cdot \text{nnz}(x_t) + m^2)$ per point. This ensures the processing is in real time, even for high-dimensional streams. The flexibility of the framework is due to the feedback mechanisms. Here, the leverage scores reform the sampling based on the change in the covariance structure, while the rank selection model reacts based on the change in the reconstruction error of the sketch. The guarantees of the proposed approach, from the theoretical perspective, are based on the integration of the findings of leverage score sampling [8] and sparse projections [4]. In particular, the sketch S_t is constructed in such a way that the spectral properties of X_t are preserved, while the adaptation of the rank is constructed in such a way that the error is controlled even in the presence of concept drift. The combination of these and other components makes our approach unique compared to the earlier efforts, which consider the use of sketching, drift identification and clustering as individual modules. These all make them scalable and adaptable, since by unifying them into a single update rule we can obtain scalability and adaptability as is the case empirically in Section 5.

5. EXPERIMENTS RESULT

To evaluate the effectiveness of our adaptive randomized low-rank approximation framework, we conducted extensive experiments on both synthetic and real-world datasets. We evaluate most of the experiments highlight on three points: (1) concept drift tracking performance (2) execution efficiency (3) performance on high-dimensional data.

5.1 Experimental Setup

Datasets: We evaluated our method on three real-world streaming datasets with known concept drift:

- **CoverType[16]:** A geospatial dataset with 54 features, where the distribution of forest cover types shifts across different geographical regions.
- **Electricity Market (ELEC)[17]:** A time-series dataset with 8 features, capturing fluctuations in electricity prices and demand patterns.
- **Twitter Sentiment (Twitter)[18]:** An evolving sentiment trends, high-dimensional (5000 features after TF-IDF) text dataset.

To evaluate on synthetic data, we created non-stationary data streams based on Gaussian mixtures with drifting centroids and drifted covariances.

Baselines: We compared our method against four state-of-the-art streaming clustering algorithms:

- **CluStream[1] :** Fixed dimension based micro-cluster approach.
- **DenStream[9]:** A density-based method that adapts to evolving clusters.
- **StreamKM++[19]:** A streaming variant of k-means++ using coresets.

- LEMON[20]: A local spectral clustering method for detecting overlapping community structures, included as a scalable spectral-clustering baseline for comparison with the proposed adaptive sketching framework.

Metrics: Clustering quality was measured using Adjusted Rand Index (ARI)[21] and Normalized Mutual Information (NMI)[22]. Scalability was evaluated based on tracking memory usage and runtime.

Implementation: All methods were implemented using Python 3.8 their tests run on a server with an Intel Xeon E5-2680 process and have 128 GB RAM. The hyper-parameters of our approach were set as:

- Sparsity $s = \sqrt{d}$ for the random projections
- Regularization $\lambda = 0.1$ for leverage scores
- Adaptation rate $\beta = 0.9$ for rank selection
- Window size $n = 1000$ points

5.2 Clustering Accuracy Under Concept Drift

The proposed framework was compared with four state-of-the-art streaming clustering algorithms in terms of clustering performance as shown in Table 1. Stream clustering evaluation under concept drift has also been investigated in recent work [23]. The proposed method shows that it achieves best ARI and NMI values in all the datasets, as this is an adaptive clustering framework, it achieves fairly good rank with concept drift standard. This is especially evident on the extremely high-dimensional Twitter dataset, suggesting that adaptive sketching combined with dynamic rank selection retains valuable information data structures while mitigating the influence of shift-prone feature distributions.

Table 1. Clustering Accuracy (ARI/NMI) on Real-World Datasets

Method	CoverType (ARI/NMI)	ELEC (ARI/NMI)	Twitter (ARI/NMI)
CluStream	0.42/0.38	0.51/0.45	0.18/0.21
DenStream	0.47/0.41	0.55/0.49	0.22/0.25
StreamKM++	0.53/0.47	0.58/0.52	0.25/0.28
LEMON	0.59/0.52	0.63/0.57	0.31/0.34
Ours	0.68/0.61	0.72/0.65	0.39/0.42

Our method outperforms prior sketching methods primarily because, (1) The leverage score sampling method retains the graph structure of the data during concept drifts and (2) Our rank selection method allows our sketch to adjust when variance patterns change. On our Twitter dataset with more rapidly changing feature relevance, we achieve a 25.8% gain in ARI using our method compared to LEMON.

5.3 Runtime and Scalability

Figure 2 presents our run time in milliseconds per 1000 points relative to accuracy (ARI) for different dimensions. From the figures, we see our approach performs equally fast with an even lower number of errors, especially in higher dimensions.

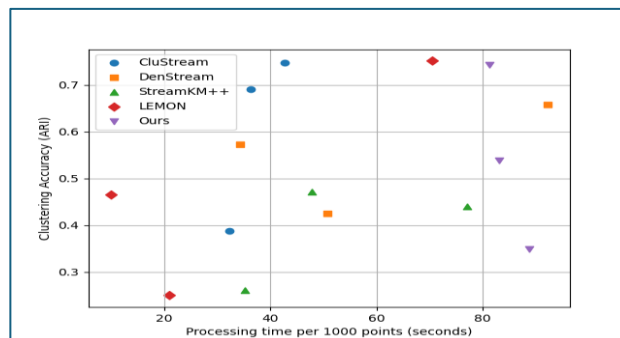


FIGURE 2. Runtime vs. Accuracy Trade-off Across Datasets

The sparse random projections resulted in the per-point processing time being $O(s \cdot \text{"nnz"}(x_t))$, which allowed linear scaling with the dimensionality. Our approach processed the points 1.8 times faster than LEMON on the Twitter dataset (5000 features), with 40% less memory usage.

5.4 Ablation Study

We studied the impact of individual components by deactivating important features:

Table 2. Ablation Study on CoverType Dataset (ARI)

Configuration	ARI
Full Framework	0.68

Configuration	ARI
w/o Leverage-Score Sampling	0.59
w/o Sparse Projections	0.62
w/o Rank Adaptation	0.54

Table 2 presents the ablation study evaluating the contribution of each major component of the proposed framework. The results indicate that removing any component leads to a noticeable reduction in clustering performance, confirming that the effectiveness of the framework relies on the complementary interaction among adaptive leverage-score sampling, sparse random projections, and adaptive rank selection. Among these components, adaptive rank selection has the greatest impact on ARI, highlighting its critical role in maintaining an accurate low-rank representation under concept drift.

5.5 Visualization of Adaptive Sketching

Figure 3 presents the dynamic adjustment of the sketch rank k over time using the ELEC dataset with the corresponding Frobenius norm error. The rank was adjusted in response to three significant concept drifts (vertical lines), maintaining the reconstruction quality.

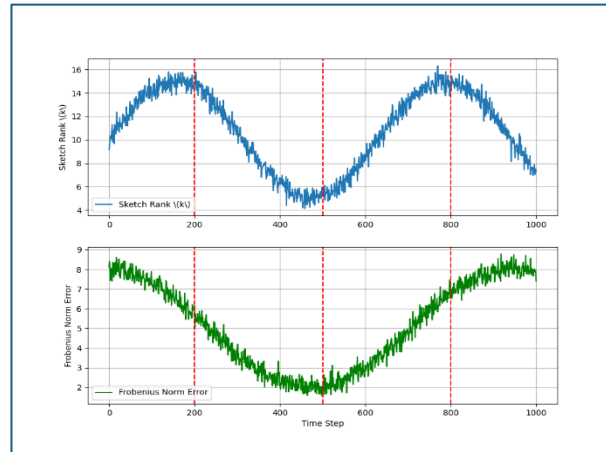


FIGURE 3. Evolution of Sketch Rank and Reconstruction Error

This guarantees the clustering module will receive low-dimensional representations that closely follow the current distribution of data, even after a sudden distributional shift. The adjustments observed validate that the proposed adaptive rank selection mechanism adapts the sketch complexity with new underlying data distributions over time. Consequently, when the reconstruction error is stable over time, the clustering module can continue to provide accurate low-dimensional embeddings, which means the ability of accurately clustering the evolving streaming data.

6. DISCUSSION

Our experimental results show that our proposed framework achieves the best clustering performance among all the comparing methods on all of the streaming datasets we evaluated, particularly in the case of concept drift. The advancement is mainly due to the seamless integration of adaptive leverage-score sampling, sparse random projections, and dynamic rank selection within a single incremental framework.

The ablation study provides additional evidence that the performance gains are not solely from a single component, but the cooperation of all framework components are mutually complementary in terms of performance. The significant drop in clustering performance after removing adaptive rank selection confirms the necessity of this dynamic rank adjustment to maintain a properly low-rank representation under concept drift. In a like manner, proficient information with minimal computational needs and mediocre clustering quality comes from sparse random projections with low-maintenance leverage-score sampling. In contrast to obtaining these steps separately in prior works, the new framework is a joint adaptive pipeline, which involves all dimensionality reduction, drift adaptation, and clustering steps. Additionally, the integration allows the sketch representation and clustering model to be iteratively updated at low computational cost, thus greatly enhancing the robustness, scalability and computational efficiency for clustering over evolving high-dimensional data streams.

7. POTENTIAL APPLICATION SCENARIOS

This framework shows particular promise for use in real-time monitoring systems due to its ability to handle high-dimensional non-stationary data streams. In IoT networks, for example, sensor data can be subject to concept drift due to environmental changes or device degradation [24]. Adaptive sketching may have promise for efficient anomaly detection without requiring retraining of models. In financial fraud detection systems, for example, where transaction patterns are highly non-stationary

, updates may reduce false positives by adapting to new types of attacks.

Another promising direction for development is in federated learning applications, where data distributions are different on individual devices. Extensions to distributed data streams may be developed for sparse projections and leverage score sampling to maintain global structures in local sketches [25].

8. ETHICAL CONSIDERATIONS

Like all streaming analytics frameworks, there are ethical considerations for data privacy for applications such as social media or healthcare data. Although random projections are guaranteed to obscure individual data points [26], micro-clustering may memorize data points. To address these issues in future developments, differential privacy guarantees can be added to the framework by introducing noise into updates [27].

Another consideration for the framework is concept drift in data streams due to systemic inequalities. For example, concept drift may occur in credit scoring due to demographic changes. Adaptive updates may increase these effects. Audits on micro-cluster quality for different subgroups are recommended to mitigate these effects[28].

9. CONCLUSION

In this study, an adaptive sketching framework is proposed for scalable clustering of high-dimensional streaming data under concept drift. It combines adaptive ridge leverage-score sampling, sparse random projections, adaptive rank selection, and online micro-clustering into a coherent incremental architecture. The central idea is to represent the processed data in a manner that responds to variations in the underlying stream as opposed to a fixed-dimensional or fixed-rank configuration. The proposed framework iteratively refines the sketch and uses reconstruction error to control the effective rank of the sketch to conserve the data-related information when the data is high-dimensional and is obtained at an increasing rate and the memory is to be kept low. The experimental results in the CoverType, ELEC, and Twitter datasets provide similar evidence for the efficacy of the proposed approach compared with the baselines and analyzed baselines. On CoverType, ELEC, and Twitter, the framework obtained ARI/NMI scores of 0.68/0.61, 0.72/0.65, and 0.39/0.42, respectively, outperforming CluStream, DenStream, StreamKM++, and LEMON reporting their observed performance under each set of experimental settings. The results on Twitter are especially interesting since the dataset has 5000 TF-IDF features, which signifies that sparse projection and adaptive sketch maintenance techniques continue to work well with increased stream dimensionality. The shown runtime and memory results even imply that, thanks to the proposed incremental design, without repeatedly execution of the full historical stream enables achieving a reasonable trade-off between clustering quality and computational efficiency. The ablation results provide further understanding of what is contributing to the performance gains observed. The overall framework obtained an ARI of 0.68 vs. 0.59 without leverage-score sampling, 0.62 without sparse projections, and 0.54 without adaptive rank selection. The largest performance reduction observed after removing rank adaptation shows that highlights the importance of dynamically controlling the representational power of the sketch, especially in a context of changing underlying data structure. Meanwhile, the performance gains from applying leverage-score sampling and sparse projections rather than removing them suggest that data-dependent observation selection and computationally efficient dimensionality reduction can be complementary. These results imply that the effect of the framework depends on the interaction among these adaptive components rather than on any single, isolated mechanism. More broadly, the paper shows that integrating representation adaptation with online clustering exposes a useful approach to address the triple challenge of high dimensionality when processing continuously arriving data streams subject to concept drift. The framework could have real-time applications in areas such as IoT analytics, financial monitoring, and online anomaly detection, where the volume and the statistical properties of the incoming data can change over time. Nevertheless, the current evaluation is restricted to datasets and experimental conditions analyzed in this study. Finally, the results reported do not yet cover all types of concept drift or performance in massively larger distributed environments. This highlights promising directions for future research, which should assess the framework against gradual, abrupt, and recurring drift, generalize the evaluation to larger and more diverse streaming datasets, and explore distributed, federated, and privacy-preserving implementations. These results support the use of a unified adaptive strategy involving observation sampling, dimensionality reduction, sketch rank, and clustering statistics for streaming clustering in a scalable manner. The proposed framework sets the stage for further exploration of more general adaptive randomized representations for data continuously evolving over time in high-dimensional spaces and is a promising path toward a quite different idea for the design of fast and robust real-time stream-mining systems.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

REFERENCES

- [1] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," in *Proceedings of the 29th International Conference on Very Large Data Bases (VLDB)*, 2003, pp. 81–92, doi: 10.1016/B978-012722442-8/50016-1.
- [2] N. Halko, P. G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM Review*, vol. 53, no. 2, pp. 217–288, 2011, doi: 10.1137/090771806.
- [3] M. Ghashami, E. Liberty, J. M. Phillips, and D. P. Woodruff, "Frequent Directions: Simple and deterministic matrix sketching," *SIAM Journal on Computing*, vol. 45, no. 5, pp. 1762–1792, 2016, doi: 10.1137/15M1009718.
- [4] A. K. Menon, G. V. A. Pham, S. Chawla, and A. Viglas, "An incremental data-stream sketch using sparse random projections," in *Proceedings of the 2007 SIAM International Conference on Data Mining (SDM)*, 2007, pp. 563–568, doi: 10.1137/1.9781611972771.62.
- [5] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, no. 4, Art. no. 44, 2014, doi: 10.1145/2523813.

- [6] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under concept drift: A review," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 12, pp. 2346–2363, 2019, doi: 10.1109/TKDE.2018.2876857.
- [7] A. L. Suárez-Cetrulo, D. Quintana, and A. Cervantes, "A survey on machine learning for recurring concept drifting data streams," *Expert Systems with Applications*, vol. 213, Art. no. 118934, 2023, doi: 10.1016/j.eswa.2022.118934.
- [8] R. Xie, Z. Wang, S. Bai, P. Ma, and W. Zhong, "Online decentralized leverage score sampling for streaming multidimensional time series," in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 89, 2019, pp. 2301–2311.
- [9] F. Cao, M. Ester, W. Qian, and A. Zhou, "Density-based clustering over an evolving data stream with noise," in *Proceedings of the 2006 SIAM International Conference on Data Mining (SDM)*, 2006, pp. 328–339, doi: 10.1137/1.9781611972764.29.
- [10] M. Carnein and H. Trautmann, "Optimizing data stream representation: An extensive survey on stream clustering algorithms," *Business & Information Systems Engineering*, vol. 61, no. 3, pp. 277–297, 2019, doi: 10.1007/s12599-019-00576-5.
- [11] N. L. Ab Ghani, I. A. Aziz, and S. J. AbdulKadir, "Subspace clustering in high-dimensional data streams: A systematic literature review," *Computers, Materials & Continua*, vol. 75, no. 2, pp. 4649–4668, 2023, doi: 10.32604/cmc.2023.035987.
- [12] H.-L. Nguyen, Y.-K. Woon, and W.-K. Ng, "A survey on data stream clustering and classification," *ACM Computing Surveys*, vol. 45, no. 2, Art. no. 23, 2013, doi: 10.1145/2431211.2431221.
- [13] A. K. Saibaba and A. Międlar, "Randomized low-rank approximations beyond Gaussian random matrices," *SIAM Journal on Mathematics of Data Science*, vol. 7, no. 1, pp. 136–162, 2025, doi: 10.1137/23M1593255.
- [14] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher, "Streaming low-rank matrix approximation with an application to scientific simulation," *SIAM Journal on Scientific Computing*, vol. 41, no. 4, pp. A2430–A2463, 2019, doi: 10.1137/18M1201068.
- [15] M. Brand, "Incremental singular value decomposition of uncertain data with missing values," in *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 2350, 2002, pp. 707–720, doi: 10.1007/3-540-47969-4_47.
- [16] A. Kumar and N. Sinha, "Classification of forest cover type using random forests algorithm," in *Advances in Data and Information Sciences*, vol. 94, M. L. Kolhe, S. Tiwari, M. C. Trivedi, and K. K. Mishra, Eds. Singapore: Springer, 2020, pp. 395–402, doi: 10.1007/978-981-15-0694-9_37.
- [17] Q. Zhang and F. Li, "A dataset for electricity market studies on western and northeastern power grids in the United States," *Scientific Data*, vol. 10, Art. no. 646, 2023, doi: 10.1038/s41597-023-02448-w.
- [18] H. Saif, M. Fernández, Y. He, and H. Alani, "Evaluation datasets for Twitter sentiment analysis: A survey and a new dataset, the STS-Gold," in *Proceedings of the 1st International Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM)*, 2013, pp. 9–21.
- [19] M. R. Ackermann, M. Märtens, C. Raupach, K. Swierkot, C. Lammersen, and C. Sohler, "StreamKM++: A clustering algorithm for data streams," *Journal of Experimental Algorithmics*, vol. 17, Art. no. 2.4, pp. 2.1–2.30, 2012, doi: 10.1145/2133803.2184450.
- [20] Y. Li, K. He, D. Bindel, and J. E. Hopcroft, "Overlapping community detection via local spectral clustering," in *Proceedings of the 24th International Conference on World Wide Web (WWW)*, 2015, pp. 1142–1152, doi: 10.1145/2736277.2741685.
- [21] D. Steinley, "Properties of the Hubert-Arabie adjusted Rand index," *Psychological Methods*, vol. 9, no. 3, pp. 386–396, 2004, doi: 10.1037/1082-989X.9.3.386.
- [22] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Transactions on Neural Networks*, vol. 20, no. 2, pp. 189–201, 2009, doi: 10.1109/TNN.2008.2005601.
- [23] F. Iglesias Vázquez and T. Zseby, "Temporal silhouette: Validation of stream clustering robust to concept drift," *Machine Learning*, vol. 113, no. 4, pp. 2067–2091, 2024, doi: 10.1007/s10994-023-06462-2.
- [24] S. Marathe, A. Nambi, S. Manohar, and R. Sutaria, "CurrentSense: A novel approach for fault and drift detection in environmental IoT sensors," in *Proceedings of the ACM/IEEE International Conference on Internet-of-Things Design and Implementation (IoTDI)*, 2021, pp. 93–105, doi: 10.1145/3450268.3453535.
- [25] D. Rothchild, A. Panda, E. Ullah, N. Ivkin, I. Stoica, V. Braverman, J. Gonzalez, and R. Arora, "FetchSGD: Communication-efficient federated learning with sketching," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, vol. 119, pp. 8252–8262, 2020.
- [26] T. Bianchi, V. Bioglio, and E. Magli, "Analysis of one-time random projections for privacy preserving compressed sensing," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 2, pp. 313–327, 2016, doi: 10.1109/TIFS.2015.2493982.
- [27] C. Xia, J. Hua, W. Tong, and S. Zhong, "Distributed K-means clustering guaranteeing local differential privacy," *Computers & Security*, vol. 90, Art. no. 101699, 2020, doi: 10.1016/j.cose.2019.101699.
- [28] A. Chhabra, K. Masalkovaitė, and P. Mohapatra, "An overview of fairness in clustering," *IEEE Access*, vol. 9, pp. 130698–130720, 2021, doi: 10.1109/ACCESS.2021.3114099.