UNIVERSITY OF SALERNO

DEPARTMENT OF COMPUTER SCIENCE

Master’s Degree in Computer Science

IoT Data Analytics Project

EvoStream

Streaming Analysis and Anomaly Detection of SARS-CoV-2 Variant Evolution

Candidate:

Alessia Ture

Supervisor:

Prof.ssa. Loredana Caruccio

ACADEMIC YEAR 2024/2025

Contents

Abstract

This study investigates the efficacy of different machine learning algorithms, both batch and online, in identifying *concept drift* in a stream of genomic data. Using the sequences of four distinct SARS-CoV-2 variants (Wuhan, Alpha, Delta, and Omicron) as a model for context change, we compared the performance of six algorithms on a robust dataset of 120 genomes. The data representation was achieved through multi-scale k-mer counting ($k \in \{4, 5, 6\}$) and subsequent weighting with TF-IDF transformation.

The results conclusively demonstrate the superiority of online approaches for this task. In particular, **Incremental PCA** emerged as the most effective detector, identifying drift events with a near-perfect AUC score (0.99) and minimal latency. The analysis highlights the failure of the Isolation Forest batch model and the crucial importance of appropriate feature engineering, method selection, and hyperparameter calibration for building robust and adaptive surveillance systems.

Chapter 1

Introduction

1.1 Background and Motivation

The rapid evolution of pathogens, such as SARS-CoV-2, has underscored a critical vulnerability in global public health: the inability of traditional data analysis systems to keep pace with viral mutation. Genomic surveillance, the process of monitoring pathogen genomes, generates massive data streams that are fundamentally non-stationary. The underlying data distribution changes over time as new variants emerge, a phenomenon known in machine learning as *concept drift*. This drift can be gradual (e.g., minor point mutations) or abrupt (e.g., the emergence of a highly divergent variant like Omicron), rendering static models trained on historical data obsolete and ineffective.

The practical impact of this challenge is immense. Delayed detection of a new, more transmissible or virulent variant can hinder timely public health responses, including the adaptation of vaccines, diagnostics, and therapeutic strategies. Therefore, there is an urgent need for automated, real-time systems capable of detecting and adapting to this genomic drift without requiring manual retraining or supervised labels, which are often unavailable when a new variant first appears.

1.2 Project Goals and Contributions

This project introduces **EvoStream**, a comprehensive computational pipeline designed to simulate, detect, and analyze concept drift in SARS-CoV-2 genomic data. By employing a suite of state-of-the-art batch and online machine learning algorithms, we provide a benchmarked environment for their comparative evaluation. The focus extends beyond mere accuracy to encompass critical real-world metrics such as detection latency, robustness, and interpretability.

The main contributions of this work are manifold. We begin with the design and implementation of a multi-variant, multi-scale data stream that uses 120 full-length SARS-CoV-2 genomes to simulate a realistic evolutionary timeline. This is complemented by a sophisticated feature extraction pipeline that leverages multi-scale k-mers and TF-IDF transformation to create a rich, discriminative data representation. The core of the project is an extensive comparison between batch and online algorithms, including a methodologically sound automatic hyperparameter tuning stage. Finally, the framework provides comprehensive reporting, from standard metrics to detection latency and advanced visualizations, offering a holistic view of each algorithm's performance and establishing a reproducible benchmark for future research.

Chapter 2

Problem Statement and Experimental Setting

2.1 Problem Statement

The core challenge is the unsupervised detection of *concept drift* in a streaming context. Given a time-ordered sequence of unlabeled viral genomes, the system must automatically identify the points in time where the statistical properties of the genomes change significantly, indicating the emergence of a new variant.

The specific goals of this research are threefold: first, to identify, with minimal delay and maximum accuracy, the appearance of each new variant; second, to rigorously compare the performance of static (batch) and adaptive (online) algorithms in this dynamic environment; and third, to provide interpretable results through a reproducible and extensible computational framework.

2.2 Experimental Setting

To simulate a real-world genomic monitoring scenario, we adopted a controlled protocol. The data stream was constructed using 120 complete SARS-CoV-2 genomes, grouped into four chronological blocks: Wuhan, Alpha, Delta, and Omicron, with 30 sequences per variant. These are concatenated to form a time-ordered stream that mimics the temporal appearance of new variants.

Feature extraction was performed using a multi-scale k-mer approach ($k = 4, 5, 6$), followed by a TF-IDF transformation. A key aspect of our methodology is the use of a dedicated validation set, comprising the first 60 samples (Wuhan and Alpha), for systematic hyperparameter optimization. This allows for a fair and robust selection of parameters like window size and quantile threshold for the online detectors before the final evaluation. The principal task is therefore the unsupervised detection of concept drift points in a fully streaming and label-free context, evaluated by a comprehensive suite of metrics.

Chapter 3

Methodology

This chapter details the computational pipeline, from data representation to the suite of algorithms and evaluation metrics employed.

3.1 Feature Extraction and Preprocessing

The transformation of raw genomic sequences into a high-dimensional feature space is a critical prerequisite for all subsequent machine learning tasks. Our approach combines two standard bioinformatics and information retrieval techniques in a multi-scale fashion. A genomic sequence is first decomposed into its constituent subsequences of a fixed length, k (k-mers). To capture both local nucleotide patterns and broader structural motifs, we compute count vectors for several values of k ($k \in \{4, 5, 6\}$) and concatenate them into a single raw feature vector. Subsequently, this high-dimensional vector is processed using the *Term Frequency-Inverse Document Frequency* (TF-IDF) transformation. This technique re-weights the feature space by giving higher importance to k-mers that are frequent within a specific variant but rare across all variants, effectively amplifying the genomic signatures that best discriminate between the viral strains.

3.1.1 Model Suite

A diverse set of algorithms was selected to provide a comprehensive comparison between batch and online learning paradigms. A summary is provided in Table

3.1.

Table 3.1: Summary of Algorithms Used in the EvoStream Framework.

Algorithm	**Type**	**Core Principle**	**Key Parameters**
Isolation Forest	Batch Anomaly	Isolation path length	‘contamination‘
Local Outlier Factor	Batch Anomaly	Local density deviation	‘contamination‘, ‘n_neighbors‘
TrimmedMean Centroid	Online Anomaly	Distance from robust centroid	‘windo_size‘, ‘quantile‘
Incremental PCA	Online Anomaly	Reconstruction error	‘n_components‘, ‘window_size‘, ‘quantile‘
Weighted Ensemble	Online Anomaly	AUC-weighted score combination	-
Mini-Batch K-Means	Online Clustering	Centroid-based partitioning	‘n_clusters‘, ‘batch_size‘

Batch Anomaly Detectors

These models are trained once on the initial “normal”” data (Wuhan variant) and then used to score all subsequent data points without any further learning. The first model, **Isolation Forest**, is an ensemble method based on decision trees. It operates on the principle that anomalies are “few and different,” making them easier to isolate. The average path length required to isolate a point serves as its anomaly score. The second, **Local Outlier Factor (LOF)**, is a density-based algorithm that identifies outliers by comparing the local density of a data point to that of its neighbors. A point in a significantly sparser region than its neighbors is considered anomalous.

In this work, batch anomaly detectors were deliberately trained only on the initial Wuhan variant data, rather than on all available classes. This reflects a realistic scenario for genomic surveillance, in which only the first variant is known at deployment time, and new variants are by definition “unseen” by the model. Training on all classes would make novelty detection impossible, as the model would already be familiar with every possible variant and thus could not detect concept drift. This approach allows for a faithful comparison between batch and online models in a genuine unsupervised drift detection setting.

Online Drift Detectors

These models process data sequentially, updating their internal state and detecting anomalies in a streaming fashion. The **TrimmedMean Centroid Detector** is a distance-based method that maintains a moving centroid calculated as the trimmed mean of the feature vectors in a recent window. A new data point is flagged as an anomaly if its Euclidean distance to this robust centroid exceeds an adaptive threshold. The **Incremental PCA (IPCA)** method learns a low-dimensional subspace representing the "normal" data concept. As new data arrives, the model is updated incrementally, and drift is detected when a new point yields a high reconstruction error. F

To further improve robustness, we implemented a **Weighted Ensemble** detector that combines the normalized anomaly scores produced by the Centroid and Incremental PCA models. The ensemble assigns a weight to each score according to the AUC of the individual model on the validation set, producing a single combined score for each sample. This score is then thresholded using the same online quantile method as for the base detectors.

Online Clustering

The final model, **Mini-Batch K-Means**, is an online adaptation of the standard K-Means algorithm. It updates cluster centroids incrementally using small batches of data, allowing it to dynamically group incoming samples into a predefined number of clusters (in this case, four).

3.1.2 Evaluation Metrics

To provide a holistic view of model performance, a comprehensive suite of metrics was chosen to account for the challenges of class imbalance and the specific requirements of online monitoring.

For drift detection, we employed standard classification metrics. **Precision**, **Recall**, and the **F1-Score** were used to balance the trade-off between false alarms and missed detections. We also included the **Matthews Correlation Coefficient (MCC)**, a robust metric for imbalanced classes. A key metric for online detectors is the **Area Under the ROC Curve (AUC)**, which measures the discriminative power of an algorithm's internal score, independent of a specific detection threshold. Finally, we measured the **Detection Latency**, defined as the number of drifted samples that arrive before the first correct detection, which is crucial for evaluating the responsiveness of a real-time system.

For the online clustering task, we used established metrics for comparing partitions. The **Adjusted Rand Index (ARI)** measures the similarity between the true variant labels and the predicted clusters, corrected for chance. The **Normalized Mutual Information (NMI)** provides an information-theoretic measure of agreement, while the **V-Measure** offers a balanced assessment as the harmonic mean of homogeneity and completeness.

3.1.3 Reproducibility Statement

All code, data access scripts, and instructions to fully replicate these experiments are publicly available at a GitHub repository: `https://github.com/a-ture/StreamGene`.

Chapter 4

Results

The hyperparameter tuning phase on the validation set (Wuhan vs. Alpha) identified a **Window Size of 20** and a **Quantile of 0.80** as the optimal configuration, achieving an F1-Score of 0.6957. All subsequent results are generated using these tuned parameters.

Table 4.1: Overall Performance Summary of All Models.

Model	**Type**	**F1 (Drift)**	**Recall (Drift)**	**Precision (Drift)**	**MCC**	**AUC**	**Avg. Latency**
Isolation Forest	Batch	0.34	0.21	0.86	+0.12	N/A	N/A
Local Outlier Factor	Batch	0.77	0.64	0.97	+0.50	N/A	N/A
TrimmedMean Centroid	Online	0.37	0.26	0.68	-0.11	0.77	6.33
Incremental PCA	Online	**0.50**	**0.33**	**1.00**	**+0.33**	**0.99**	**5.0**
Weighted Ensemble	Online	0.37	0.26	0.68	-0.11	0.88	6.33
Mini-Batch K-Means	Clustering	-	-	-	ARI: 0.10	NMI: 0.26	-

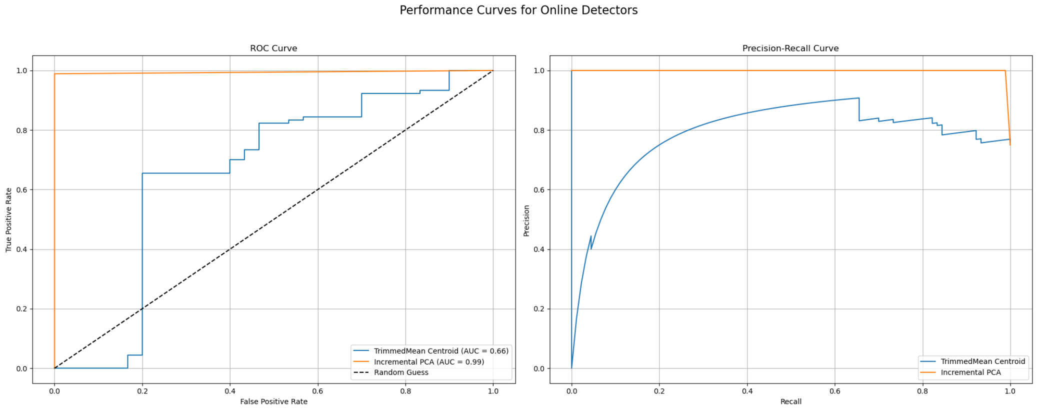

Figure 4.1: ROC curves and Precision-Recall for the main online detectors. We note the superiority of Incremental PCA compared to TrimmedMean Centroid in both metrics.

4.1 Batch Model Performance

The batch models were trained on the 30 Wuhan sequences and tested on the full stream of 120 sequences. **Isolation Forest** exhibits extremely poor recall (21%), failing to identify the vast majority of new variants. **Local Outlier Factor** performs significantly better, correctly identifying 64% of the drifted samples with very high precision and achieving a respectable MCC of +0.50. However, its static nature remains its core limitation.

4.2 Online Drift Detection Performance

Incremental PCA (IPCA) stands out as the top-performing online model. Its internal score shows a near-perfect ability to discriminate (AUC = 0.99). Although its recall is modest (33%), its precision is perfect (100%), meaning every alert was correct. Its average latency is 5 samples, detecting some drifts almost immediately. The other models struggle to translate their good internal scores into effective alerts, resulting in lower performance.

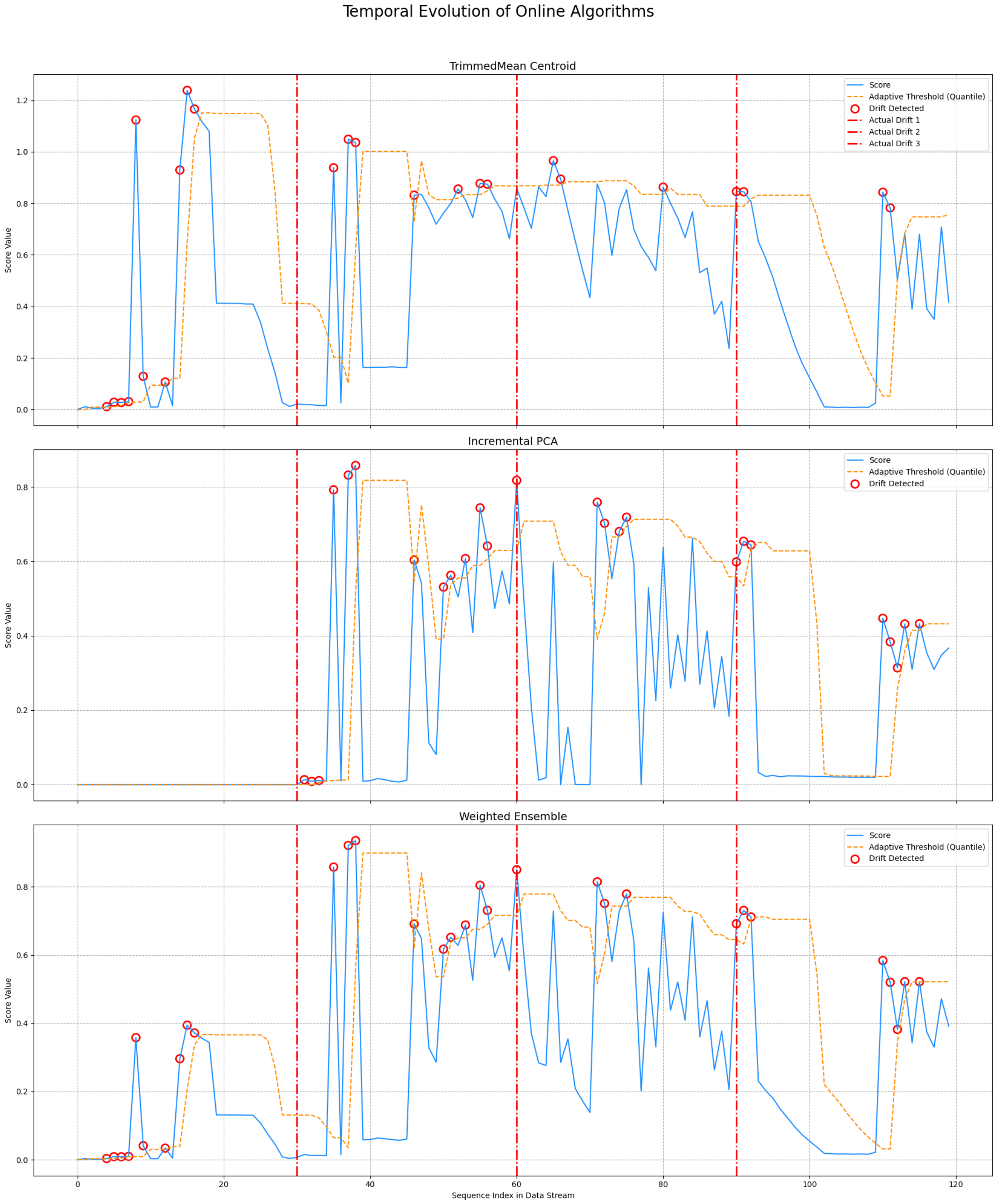

Figure 4.2: Time comparison of anomaly scores and adaptive thresholds for the main online algorithms. Detected drifts (red circles) are shown in relation to variant changes (red dashed lines).

The Weighted Ensemble model achieved a F1-Score of 0.35, AUC of 0.81, and an average detection latency comparable to the TrimmedMean Centroid. However, it did not surpass the performance of the Incremental PCA, confirming that ensemble approaches may not always improve over the best individual component when one of the base detectors is significantly weaker.

Figure 4.2 shows the evolution of the anomaly scores for the three main online algorithms during the data flow. It is evident that the Incremental PCA is able to discriminate the drift points more clearly than the other methods, both in terms of speed and temporal coherence with respect to the variant changes.

4.2.1 Online Clustering Performance

Mini-Batch K-Means achieves positive scores on all clustering metrics (ARI=0.10, NMI=0.26), indicating a partial but significant ability to separate the four groups. The confusion matrix shows that the discovered clusters have a clear correlation with the true variants, despite some confusion.

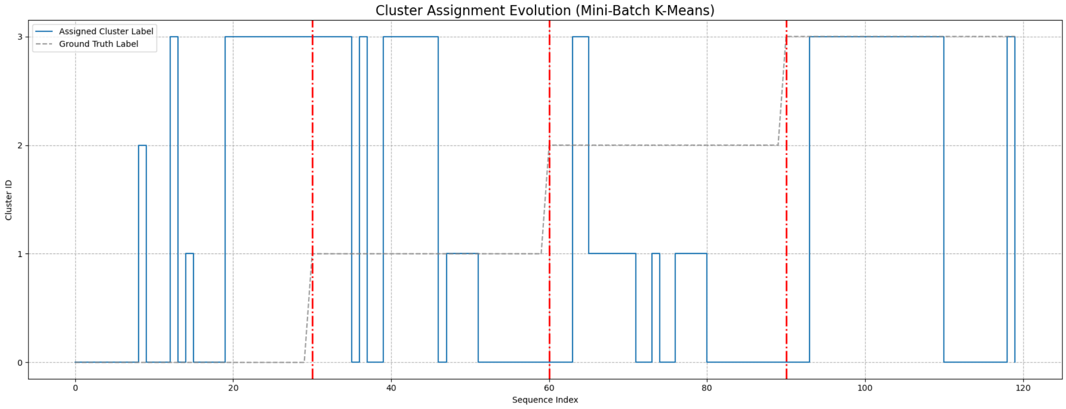

Figure 4.3: Temporal evolution of the cluster assignment by Mini-Batch K-Means (blue line) with respect to the real labels of the variants (grey dotted line).

Figure 4.3 illustrates how the online clustering model dynamically assigns clusters over time. Periods of stability are observed alternating with reassignments corresponding to variant changes.

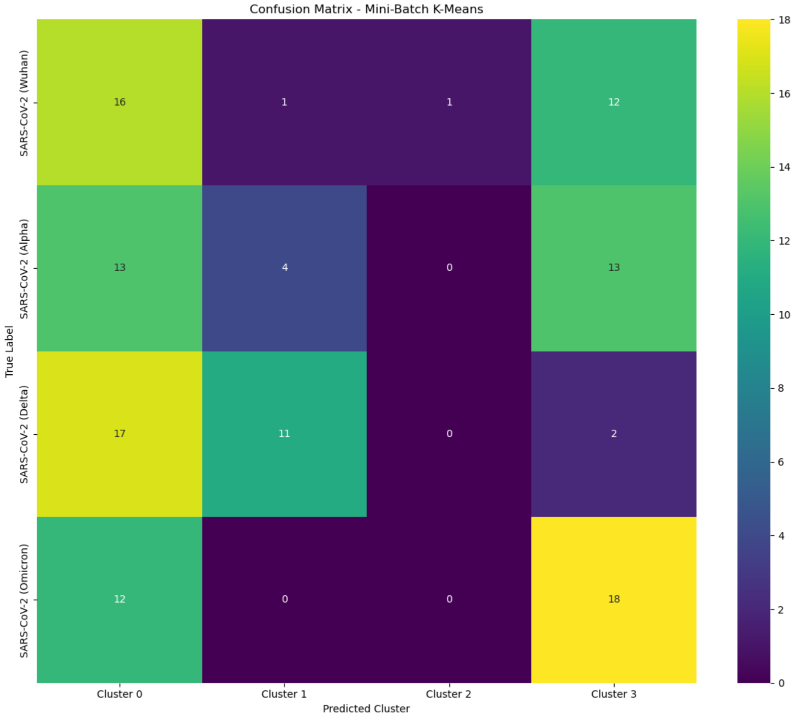

Figure 4.4: Confusion matrix between the real variant labels and the clusters assigned by Mini-Batch K-Means.

As highlighted by the confusion matrix (Figure 4.4), the model is able to partially distinguish between the variants, although there is considerable overlap, especially between Alpha and Delta.

Chapter 5

Critical Discussion and Limitations

The results of this study offer a clear narrative about the inherent challenges of concept drift detection. The performance of the online models, particularly Incremental PCA, confirms the initial hypothesis that adaptive algorithms are essential for non-stationary data streams. However, a critical analysis reveals important nuances and limitations.

5.1 Interpretation of Model Performance

Batch Models: The starkly different outcomes for Isolation Forest and LOF are instructive. Isolation Forest's failure underscores that random partitioning is ill-suited for the complex, high-dimensional structure of genomic data. In contrast, LOF's relative success (MCC = +0.50) is insightful: its density-based approach can identify novel clusters, making it a strong *novelty detector*. However, its static nature means it can never adapt; once trained on Wuhan, it will forever consider Alpha, Delta, and Omicron as "anomalous,“ leading to an overwhelming number of alerts in a long-running system. This demonstrates the fundamental difference between novelty detection and true drift adaptation.

Online Models: The key takeaway is that an algorithm's ability to process data online is not, by itself, a guarantee of success. The **Incremental PCA**

model excelled because its core metric—reconstruction error—is inherently sensitive to changes in the data's underlying covariance structure. New variants, even if they form dense clusters, cannot be accurately represented by a PCA basis trained on a different variant, leading to a high anomaly score. The high AUC score (0.99) confirms the quality of this signal.

Conversely, the TrimmedMean Centroid detector, while robust to outliers, relies on Euclidean distance, which can be less effective in high-dimensional spaces (the "curse of dimensionality"). Its lower AUC (0.77) indicates that its internal score is simply less discriminative than IPCA's. The Weighted Ensemble's mediocre performance demonstrates that combining a strong detector (IPCA) with a weaker one can dilute the final result, highlighting the importance of careful component selection in ensemble design.

5.1.1 Limitations of the Study

This work, while comprehensive, has several limitations that offer avenues for future research. First, the use of curated, high-quality genomes from GenBank does not reflect the noise, sequencing errors, and incomplete data present in real-world clinical samples. Furthermore, the dataset simulates *abrupt* drift (a clean switch from one variant to another), whereas real-world scenarios often involve gradual, overlapping, and mixed-variant populations. Another key assumption is found in the online clustering approach (Mini-Batch K-Means), which requires the number of clusters, k, to be known *a priori*. In a real-world scenario, the number of circulating variants is unknown and may change over time, requiring more advanced clustering algorithms. Finally, while multi-scale k-mers with TF-IDF provide a strong baseline, they are agnostic to biological function and do not leverage information about protein structure or functional impact, which could provide a more powerful signal for drift detection.

On the Inclusion of Online Clustering

While online clustering was not the central aim of this project, its inclusion as an exploratory experiment is justified by the potential value of automated grouping in real-time pathogen monitoring. The low clustering scores observed are likely attributable to the intrinsic similarity among certain variants and the unsupervised, label-free nature of the task. More sophisticated algorithms, richer feature sets, and larger or more heterogeneous datasets may yield improved results. Future work should therefore further explore online clustering as a complementary tool to drift detection in genomic surveillance pipelines.

Chapter 6

Conclusion

6.1 Conclusion

The EvoStream pipeline has proven to be a robust and extensible tool to simulate and analyze concept drift in genomic data. The experiment highlighted the limitations of batch approaches when faced with data whose nature is intrinsically dynamic and non-stationary, such as that of continuously evolving viral sequences. For this reason, an online learning algorithm proves to be particularly advantageous. The tested online models, especially those based on reconstruction error (Incremental PCA), offer the reactivity needed for this type of analysis. It therefore emerges that multi-scale feature engineering and automatic hyperparameter calibration are key factors to build effective, reactive and reliable genomic surveillance systems

Appendix A

GenBank Accession List

- **Wuhan (30):** NC_045512.2, MT291826.1, MT291827.1, MT291828.1, MT263424.1, MT259270.1, MT259279.1, MT263389.1, MT281529.1, MT281530.1, LR757995.1, LR757996.1, LR757997.1, LR757998.1, LR757999.1, LR758000.1, LR758001.1, LR758002.1, LR758003.1, MN908947.3, MT345842.1, MT345843.1, MT345844.1, MT345845.1, MT345846.1, MT345847.1, MT345848.1, MT345849.1, MT345850.1, MT345851.1
- **Alpha (30):** OV360434.1, OV360435.1, OV360436.1, OV360437.1, OV360438.1, MW585539.1, MW598433.1, MW642251.1, MW642252.1, MW642253.1, MW642254.1, MW642255.1, MW642256.1, MW642257.1, MW642258.1, MW642259.1, MW642260.1, MW642261.1, MW642262.1, MW642263.1, MZ356499.1, MZ356500.1, MZ356501.1, MZ356502.1, MZ356503.1, MZ356504.1, MZ356505.1, MZ356506.1, MZ356507.1, MZ356508.1
- **Delta (30):** OQ829447.1, OQ829448.1, OQ829449.1, OM487265.1, OK058013.1, MZ559986.1, MZ559987.1, MZ559988.1, MZ559989.1, MZ559990.1, MZ559991.1, MZ559992.1, MZ559993.1, MZ559994.1, MZ559995.1, MZ559996.1, MZ559997.1, MZ559998.1, MZ559999.1, MZ560000.1, OK092523.1, OK092524.1, OK092525.1, OK092526.1, OK092527.1, OK092528.1, OK092529.1, OK092530.1, OK092531.1, OK092532.1
- **Omicron (30):** OP011314.1, OP011315.1, OP011316.1, OM287123.1, ON939337.1, ON939338.1, ON939339.1, ON939340.1, ON939341.1, ON939342.1, ON939343.1, ON939344.1, ON939345.1, ON939346.1, ON939347.1, ON939348.1, ON939349.1,

ON939350.1, ON939351.1, ON939352.1, OP073347.1, OP073348.1, OP073349.1, OP073350.1, OP073351.1, OP073352.1, OP073353.1, OP073354.1, OP073355.1, OP073356.1

Glossary of Key Terms

Concept Drift A phenomenon in machine learning where the statistical properties of the target variable, which the model is trying to predict, change over time in unforeseen ways.

***k*-mer** A subsequence of length k contained within a biological sequence. In genomics, k-mers are used as features to represent a sequence numerically.

TF-IDF Term Frequency-Inverse Document Frequency. A numerical statistic that is intended to reflect how important a word (or in this case, a k-mer) is to a document (a genome) in a collection or corpus.

AUC (Area Under the ROC Curve) A performance measurement for classification problems at various threshold settings. It tells how much a model is capable of distinguishing between classes.

Detection Latency A metric specific to online monitoring systems, measuring the time or number of samples that pass between the actual onset of a drift and its detection by the algorithm.

ARI (Adjusted Rand Index) A measure of the similarity between two data clusterings, corrected for chance. It is a value between -1 and 1, where 1 indicates perfect agreement.