ADVANCED UNSUPERVISED ANALYSIS OF SPATIO-TEMPORAL PATTERN ANALYTICS

PAVLY SALEH

Thesis submitted to the University of Ottawa
in partial Fulfillment of the requirements for the
Master of Applied Science Electrical and Computer Engineering

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

Abstract

This thesis introduces a unified, scalable framework for unsupervised and annotation-efficient analysis of maritime spatio-temporal data using Automatic Identification System trajectories. Addressing challenges of manual feature engineering and label scarcity, the proposed system integrates transformer-based temporal embeddings, constrained clustering, and Active Learning. The time-series transformer is leveraged to encode vessel movements, enabling clustering to reveal latent behavioural patterns. These patterns are aggregated into bag-of-words features for track-level classification. Geospatial validation and classification accuracy confirm the semantic coherence of the learned representations. A hybrid Active Learning strategy, combining uncertainty and diversity sampling, selectively annotates sequences, accelerating model convergence while reducing required annotations by over 30%. The system addresses real-world challenges like data sparsity and class imbalance, demonstrating robustness and offering a generalizable blueprint for spatio-temporal analytics in maritime and other domains, such as aviation, logistics, and environmental monitoring, paving the way for interactive, human-in-the-loop decision-support systems in dynamic operational settings.

Cette thèse propose un cadre unifié et extensible pour l’analyse non supervisée et efficace des données spatio-temporelles maritimes via les trajectoires du Système d'Identification Automatique. Pour surmonter l’ingénierie manuelle des caractéristiques et la rareté des étiquettes, le système combine des enchâssements temporels par transformateurs, un regroupement contraint et l’apprentissage actif. Le transformateur encode les mouvements des navires, révélant des schémas comportementaux latents. Ces modèles sont ensuite regroupés en caractéristiques de type sac-de-mots pour la classification des voies. La validation géospatiale et la précision de la classification démontrent la cohérence des représentations. Une stratégie hybride d’apprentissage actif, combinant incertitude et diversité, permet une annotation sélective des séquences, accélérant la convergence et réduisant les annotations de plus de 30%. Ce système, robuste face à la rareté des données et le déséquilibre des classes, offre un modèle généralisable pour l’analyse spatio-temporelle dans le maritime, l’aviation, la logistique ou la surveillance environnementale.

Acknowledgements

First and foremost, I would like to sincerely thank Larus Technologies for their generous sponsorship of this thesis and for facilitating the infrastructure and resources that made this research possible. Their support was instrumental in ensuring that I could complete this work successfully.

I am very thankful to my supervisors, Dr. Emil Petriu and Dr. Rami Abielmona, for their invaluable advice, insight, and support throughout my entire master's journey. Their encouragement and expertise have been essential to help me achieve my academic goals, and their constructive feedback helped shape the direction of my research.

I would also like to thank my managers, Dr. John Armitage and Dr. Phillip Curtis, for their mentorship and technical guidance. Their advice and support were needed to help me overcome various obstacles and refining the practical aspects of my work.

I would also like to extend my gratitude to my friends, whose support and understanding were invaluable throughout this journey. Whether it was through lighthearted distractions, thoughtful advice, or just being there to listen, you all helped me maintain my balance when things felt overwhelming.

A special thank you goes to my parents and sister, for their love, encouragement, and unwavering support. They have been my rock throughout this process, offering encouragement and understanding every step of the way.

To my fiancée, thank you for being my anchor during this process. Your constant encouragement, positivity, and the comfort you provided during the toughest moments helped me keep going. I am incredibly grateful for your patience and love.

Finally, I would like to thank my 2 dogs Bino and Dobby, for bringing light into my life through the most stressful times. Their playful energy and unconditional affection never failed to make me smile when things got tough.

Thank you all for being a part of this journey!

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Definition
ADS-B	Automatic Dependent Surveillance–Broadcast
AIS	Automatic Identification System
AL	Active Learning
BoW	Bag-of-Words
Chronos	Transformer-based model for time-series embedding
COG	Course Over Ground (AIS feature)
Elasticsearch	Scalable search engine used as a vector database for embedding indexing
Entropy	Measure of prediction uncertainty in probabilistic classification
GPU	Graphics Processing Unit
IoT	Internet of Things
k	Number of clusters
kNN	k-nearest neighbor
KMeansConstrained	K-Means clustering with cluster size constraints
ML	Machine Learning
NLP	Natural Language Processing
ReLU	Rectified Linear Unit (activation function)
RF	Random Forest (classifier)
ROT	Rate of Turn (AIS feature)
Softmax	Activation function used in the output layer of classification models
SOG	Speed Over Ground (AIS feature)
t-SNE	t-distributed Stochastic Neighbour Embedding (dimensionality reduction)
UMAP	Uniform Manifold Approximation and Projection (dimensionality reduction)

Chapter 1

Introduction

The automated analysis of spatio-temporal patterns, data indexed by both space and time, has emerged as a critical enabler across a range of data-intensive domains including maritime surveillance, air traffic management, logistics, and sensor-driven IoT systems. As global-scale data streams become increasingly dense, dynamic, and heterogeneous, traditional rule-based approaches and handcrafted feature engineering prove inadequate for extracting actionable insights at scale. With the increasing availability of high-resolution tracking data streams, such as Automatic Identification System (AIS) for maritime vessels and Automatic Dependent Surveillance–Broadcast (ADS-B) for aircraft, the ability to efficiently analyze and interpret patterns in such data is both an operational necessity and a research challenge. The growing complexity of decision-making systems necessitates methodologies that can uncover latent behavioural structures from large, unlabelled time-series datasets, structures that are often deeply embedded within temporal sequences and exhibit significant variability across space and time [1][2].

In maritime domains, AIS data represents a rich source of vessel trajectory information, capturing vessel positions, velocities, headings, and operational states at fine temporal resolutions. Historically, maritime analytics has relied on domain-specific heuristics and manually designed features to classify vessel behaviour, detect anomalies, and model operational efficiency. However, such techniques are labour-intensive, brittle across geographic and operational contexts, and incapable of generalizing to new behaviour patterns without significant re-engineering. Furthermore, the high cost and limited availability of labelled data constrain the applicability of supervised learning paradigms, creating a pressing need for annotation-efficient, scalable, and unsupervised or semi-supervised analytical frameworks. While this data is rich in potential insight, its utility is often constrained by a lack of reliable labels and the challenges associated with manual annotation. Expert labelling of behavioural categories, such as open sea navigation, port operations, or riverine transit, is costly, inconsistent, and non-scalable. This limitation hinders the development of Machine Learning (ML)-based models for anomaly detection, classification, and predictive analytics in real-world maritime operations [3][4].

Practical use of AIS also requires nontrivial preprocessing. In this work, data cleaning involved several standard but necessary steps, including removal of duplicate or corrupted messages, temporal alignment of irregularly sampled points, coordinate smoothing, and imputation for short gaps in the signal. Numerical gaps (for example in speed or heading) were treated using local interpolation (spline or rolling-window mean where appropriate), and temporal outliers were identified and filtered using domain-aware thresholds. These preprocessing steps, described in more detail in Chapter 3, were important to ensure that embeddings encoded genuine vessel dynamics rather than artefacts of noisy transmission or intermittent coverage.

This thesis addresses these challenges through the development and validation of two interrelated, state-of-the-art methodologies that advance the frontiers of unsupervised and weakly supervised spatio-temporal analysis. The first contribution is an automated framework for maritime decision-making that leverages time-series transformer-based embeddings, scalable clustering, and vector database indexing to uncover interpretable behavioural patterns in AIS data. The second contribution builds on this foundation by integrating Active Learning (AL) into the analytical pipeline, introducing a sampling-driven approach to selectively label the most informative data instances, thereby accelerating model convergence while significantly reducing annotation overhead.

Time-series transformers play a central role in this work. The model family is built around self-attention, which learns pairwise relationships across timesteps and allows parallel processing of long sequences without the vanishing-gradient issues common in recurrent networks. In particular, this thesis uses Chronos, a transformer-style architecture adapted for multivariate time-series [11]. Chronos leverages multi-head self-attention and temporal positional encodings to capture both short- and long-range dependencies in irregular, multi-dimensional sequences. It was selected because its pretraining objectives and encoder design are well suited to multi-scale temporal correlation learning, a property that is important for vessel behaviour where events unfold over minutes to many hours. The pipeline is modular, however, and alternative time-series transformer variants can be substituted where required; Chapter 3 discusses the reasons for selecting Chronos in more detail and summarises comparative considerations.

Both frameworks are unified by their use of transformer-based temporal embeddings, specifically adapted to the unique demands of time-series data, and their emphasis on scalable, modular architectures. By segmenting AIS trajectories into 12-hour temporal windows, the systems capture operational transitions such as port entry, coastal navigation, and open-sea transit. This segmentation length was chosen after empirical exploration: shorter windows (for example 3 to 6 hours) tended to fragment coherent operational episodes, reducing cluster separability, while substantially longer windows risked mixing distinct behaviours within a single segment. The 12-hour window therefore represents a pragmatic trade-off between contextual completeness and segment-level homogeneity. Chapter 3 presents sensitivity analysis of segmentation length and discusses how different window sizes influence embedding quality and downstream clustering.

Segments are encoded using deep temporal transformers into dense vector representations that preserve spatial-temporal semantics. Embeddings are then indexed within a vector database (Elasticsearch [5]), supporting scalable clustering, anomaly detection, and similarity-based querying. The use of a vector store also simplifies incorporation of auxiliary context, because metadata and additional feature vectors can be stored and filtered alongside embeddings to support hybrid search and cluster-aware sampling.

These frameworks are explored and validated in two peer-reviewed research studies. The first, titled "Navigating Complexity: Automating Maritime Decision-Making with Temporal Transformer-Based Embeddings and Scalable Clustering" [6], is published in the 19th IEEE International Systems Conference. It demonstrates the feasibility and robustness of automated feature extraction and unsupervised pattern discovery. It achieves clustering coherence validated through silhouette analysis, geospatial alignment, and vessel-type distributions. In addition to geospatial checks, cluster validity was assessed via embedding-space visualisation (UMAP), classical clustering indices, and downstream supervised evaluation: a Random Forest classifier trained on cluster-derived labels attained a weighted F1-score of 0.90, which provides quantitative confirmation that the embedding space carries semantically meaningful signal beyond simple spatial coincidence.

The second system, developed in "Navigating the Annotation Bottleneck: Active Learning for Scalable Maritime Data Analytics" [7], published in OCEANS 2025 Great Lakes, introduces a novel AL framework designed to address the annotation bottleneck in time-series classification. This framework evaluates four AL strategies: random, uncertainty-based, diversity-based, and a hybrid uncertainty-diversity sampling approach. Applied to the transformer-derived embedding space of 5,000 AIS tracks, the hybrid strategy demonstrates superior annotation efficiency, achieving a Macro F1-score of 0.736 while reducing required labels by approximately 40% compared to random sampling. By selecting the most uncertain sample from each behaviourally coherent cluster, this strategy balances exploration and exploitation, principles fundamental to efficient learning. Experimental results are further validated through 100 repeated trials, incorporating metrics such as macro F1, macro precision, weighted accuracy, and model uncertainty entropy.

A number of methodological alternatives and validation metrics were considered during development. Density-based clustering approaches such as DBSCAN or HDBSCAN were explored, since they can adapt to varying local densities and detect arbitrarily shaped clusters; however, for large-scale operational datasets they proved sensitive to parameter choice and uneven sampling density, which in turn complicated downstream Active Learning. K-Means, in its constrained variant, was chosen because it provides scalable, reproducible clusters with enforceable minimum sizes, a desirable property when designing cluster-aware sampling strategies. In addition to cluster purity, other external evaluation measures are used to triangulate quality, including silhouette score, Adjusted Rand Index and classifier-based metrics on hold-out data; these complementary metrics are discussed in Chapters 3 and 5. Probabilistic techniques, for example hidden Markov models, Gaussian mixture models, and Bayesian nonparametric approaches, were also considered. While such methods offer principled uncertainty estimates and temporal modelling in some settings, they did not scale as effectively to high-dimensional transformer embeddings in preliminary experiments, nor did they capture long-range temporal dependencies as compactly as self-attention encoders. Nonetheless, probabilistic approaches remain a valuable avenue for future work, particularly for explicit uncertainty quantification and model calibration in operational deployments.

The thesis explicitly recognises the consequences of misclassification in operational contexts. In a research setting, misclassification primarily affects interpretive fidelity and the effectiveness of downstream analytics: for example, annotating a transit as loitering could bias risk assessments or generate false alerts. In practical deployments, misclassification can translate into operational costs such as unnecessary inspections, misallocation of monitoring resources, or failure to detect genuinely anomalous behaviour. To mitigate these risks, the pipeline adopts several safeguards: cluster-based interpretability to support human review, Active Learning to focus labelling on ambiguous instances, and multiple evaluation metrics to monitor both overall performance and per-class performance for minority behaviours. Cost-sensitive loss functions and calibrated probability outputs are recommended as future enhancements for deployments where the asymmetric cost of different error types can be quantified.

Class imbalance was an explicit characteristic of the experimental dataset and of real-world AIS streams more generally, with dominant classes such as open-sea transit far outnumbering rarer behaviours. The study deliberately preserved this imbalance to ensure ecological validity, and the Active Learning framework was designed to be robust under these conditions. In particular, the hybrid sampling method maintains behavioural coverage via cluster-aware sampling while targeting high-uncertainty examples for labelling, which improves minority-class exposure without excessive oversampling of majority behaviour. Chapter 5 provides detailed class-distribution statistics and per-class results.

Finally, the architecture is intentionally modular to support multi-modal extensions. Although this thesis focuses on AIS-derived motion data, the vector database and transformer-based embedding pipeline accommodate auxiliary inputs, for example environmental data (wind, wave, visibility), radar or satellite imagery, and port metadata. Multi-modal fusion could be realised through cross-attention mechanisms or late fusion of modality-specific embeddings, enabling richer contextual representations and potentially improving anomaly detection and forecasting performance; these directions are discussed in Chapter 6.

Taken together, these contributions form a comprehensive approach to spatio-temporal pattern analytics that is robust to data sparsity, scalable across operational environments, and adaptable to new data modalities. The methodological integration of time-series transformers, vector database infrastructure, and Active Learning constitutes a significant advancement in the field, offering a practical, interpretable, and annotation-efficient pathway toward intelligent decision-support systems in dynamic environments.

This thesis proceeds as follows: Chapter 2 provides a comprehensive review of the literature, covering foundational work in temporal representation learning, clustering, vector search systems, and annotation-efficient learning paradigms. Chapter 3 introduces the unsupervised embedding and clustering pipeline, detailing transformer model architecture, preprocessing routines, clustering validation, and geospatial consistency checks. Chapter 4 presents the Active Learning extension, covering sampling strategies, neural model architecture, iterative training pipeline, evaluation metrics, and empirical comparisons across 100 experimental repetitions. Chapter 5 presents the experimental validation, performance evaluation, empirical comparison across methods, and discusses limitations (e.g., class imbalance, temporal ambiguity). Chapter 6 concludes with a synthesis of contributions, discussion of system limitations, and directions for future research, including applications in other domains and potential for multi-modal integration and predictive extensions.

This thesis makes the following key contributions:

- **Temporal Embedding of Vessel Behaviour:** Introduces the application of a time-series transformer-based model, for encoding 12-hour AIS segments into dense embeddings that capture long-range behavioural dependencies without relying on hand-engineered features.

- **Unsupervised Clustering with Behavioural Interpretability:** Develops a two-stage clustering pipeline using K-Means to ensure both fine-grained behavioural granularity and interpretable coarse grouping. Clusters are validated using silhouette metrics, classification performance, and geospatial overlays.

- **Bag-of-Words Representation of Trajectories:** Proposes a novel bag-of-words (BoW) embedding scheme that aggregates cluster-labelled segments into trajectory-level summaries, facilitating interpretable classification of complex vessel behaviours.

- **Active Learning Framework with Cluster-Aware Sampling:** Designs and evaluates four sampling strategies within an iterative Active Learning loop. A hybrid strategy, selecting the most uncertain sample per cluster, outperforms conventional methods in both convergence speed and performance stability.

- **Empirical Validation on Real-World Maritime Data:** Demonstrates the effectiveness of the pipeline on 5,000 real-world AIS tracks, using 100-trial experimental repetitions to confirm statistical robustness and model generalisability under severe class imbalance.

By bridging the gap between unsupervised representation learning and interactive label-efficient supervision, this thesis advances the state of the art in spatio-temporal data analytics and sets a foundation for real-time, scalable, and human-in-the-loop systems capable of adapting to evolving operational contexts.

Chapter 2

Literature Review

This chapter reviews the foundational and contemporary literature underlying the analytical framework proposed in this thesis. It is structured into five core themes: (1) maritime spatio-temporal data and behavioural analytics, (2) temporal representation learning via transformers, (3) clustering for behavioural pattern discovery, (4) vector databases for scalable embedding management, and (5) Active Learning for annotation efficiency. Together, these areas of research converge to support the development of scalable, interpretable, and annotation-efficient systems for maritime and other spatio-temporal domains.

2.1 Maritime Spatio-Temporal Behavioural Analytics

2.1.1 AIS Data and Traditional Methods

Automatic Identification System (AIS) data has long served as the foundation for maritime situational awareness and anomaly detection. It provides high-frequency positional, speed, heading, and identification data at high temporal granularity. Traditionally, maritime behavioural analysis relied on domain-specific heuristics and handcrafted features to detect anomalies, define navigation phases, and classify operational zones [8]. These handcrafted features included speed over ground (SOG), rate of turn (ROT), and distance to port, which were often combined with rule-based models or clustering techniques such as DBSCAN [9].

However, the rigidity and low generalizability of handcrafted methods have become apparent with the expansion of global AIS coverage. These methods typically struggle in heterogeneous or dynamic operational contexts, requiring labour-intensive reconfiguration. For instance, speed thresholds that indicate loitering in open-sea traffic may not apply to constrained river environments. Furthermore, most traditional systems lack the capacity to capture the temporal evolution of behaviours, treating trajectory segments as isolated snapshots rather than dynamic sequences. This limitation has been highlighted in works such as those by Pallotta et al. [2], where unsupervised frameworks still required significant manual feature engineering tailored to context-specific parameters. Recent critiques emphasize that static thresholds for metrics like SOG fail to account for vessel-specific behaviour patterns (e.g., cargo ships vs. fishing vessels) or environmental factors such as currents and weather [9].

Practical use of AIS also requires substantial preprocessing. Raw AIS feeds commonly contain duplicate or corrupted messages, variable transmission intervals, missing values and brief outages, spurious locations due to GPS noise, and inconsistent static metadata (for example vessel type or MMSI reuse). Addressing these issues typically requires a pipeline of checks and transforms: removal of duplicates and obviously invalid fixes, temporal alignment of irregularly sampled points, smoothing of high-frequency jitter (for example via moving-window filters), and interpolation over short gaps using spline or local-mean methods. In some contexts, simple map-matching or proximity-to-coastline heuristics are applied to identify implausible fixes. These cleaning steps are important because noisy inputs can dominate learned representations and mislead downstream clustering or sampling strategies; Chapter 3 documents the precise cleaning choices used in this work and their effect on embedding stability.

Recent studies have proposed data-driven alternatives, including trajectory similarity metrics [3], semantic compression [4], and statistical learning for pattern extraction. Nonetheless, these methods often falter at scale or suffer from interpretability challenges, particularly when applied to raw, unfiltered data containing behavioural ambiguity or noise. For example, trajectory similarity metrics like Dynamic Time Warping (DTW) struggle with computational complexity when applied to large AIS datasets, while semantic compression techniques risk oversimplifying nuanced behaviours like "illegal transshipment" into generic "suspicious activity" labels.

2.1.2 Label Scarcity and Class Imbalance

One of the most persistent challenges in maritime analysis is the scarcity of labelled behavioural data. Given the expert knowledge required to interpret vessel behaviour in varying maritime contexts, annotating large-scale AIS datasets is both costly and time-consuming. Labels for behaviours such as anchoring, near-shore loitering, or riverine transit are often missing, inconsistently defined, or highly subjective [4]. For instance, labelling "illegal fishing" requires domain expertise to distinguish from legal trawling, often necessitating collaboration with maritime enforcement agencies.

This issue is compounded by significant class imbalance in real-world AIS streams. Dominant operational behaviours such as open-sea transit occur far more frequently than specialized behaviours like port manoeuvring or inland navigation. As highlighted in recent studies [6], supervised learning models trained on such imbalanced datasets exhibit poor generalization to underrepresented classes. This skewed distribution also affects unsupervised learning, where clustering algorithms may prioritise dominant behaviours and suppress rare, yet operationally critical, modes.

Recent works such as in [4] have proposed semi-supervised and anomaly detection-based approaches to circumvent labelling requirements, but these still struggle with the semantic labelling and interpretation of output clusters. The semi-supervised approaches attempt to mitigate this by propagating labels from a small annotated subset, but performance degrades when rare behaviours (e.g., "dark activity" where vessels disable AIS) lack even minimal labelled examples. The lack of consistent ground truth remains a central obstacle in maritime analytics. Weak supervision techniques, such as heuristics based on geofencing, have shown promise but introduce label noise that complicates model training [6].

2.2 Temporal Representation Learning with Transformers

2.2.1 From Handcrafted Features to Learned Representations

The evolution from manually engineered features to automatically learned representations marks a turning point in time-series analytics. Transformer-based models, originally designed for textual sequences (e.g., BERT [10]), have emerged as powerful alternatives for learning semantic structure from sequential data. Unlike recurrent networks (e.g., LSTMs), transformers use self-attention mechanisms to capture relationships across the entire sequence without explicit recurrence, allowing for parallelization and long-range dependency modelling.

This architectural innovation has translated effectively into time-series domains. Chronos [11], a transformer architecture specifically adapted for multivariate time-series, has demonstrated effectiveness across healthcare, sensor data, and mobility tracking. Chronos and similar time-series transformers operate by tokenizing or windowing the raw time-series into segments, applying temporal positional encodings so that order information is preserved, and using multi-head self-attention to weight contributions from different times and features. Such models can learn multiscale temporal patterns, for example a slow deceleration trend spanning hours and a rapid heading change spanning minutes, within a single representation.

There are several practical considerations when adopting time-series transformers for AIS. First, tokenization and window size interact strongly with the downstream task: very short tokens may fragment coherent behaviours, while very long tokens can mix distinct behaviours and increase computational burden. Second, transformers are computationally intensive, both in training and inference, and may require efficiency techniques such as gradient checkpointing, parameter sharing, or distillation for real-time operation. Third, transformer embeddings are high dimensional and may necessitate dimensionality reduction or approximate nearest neighbour indexing when deployed at scale. Despite these challenges, learned embeddings mitigate many brittleness issues of handcrafted features and enable scalable downstream tasks such as clustering, anomaly detection, and classification.

Because transformer design is an active area of research, alternative architectures and pretraining regimes exist, each with trade-offs in computation, receptive field, and inductive bias. Chronos was selected in this thesis for its suitability to multiscale temporal correlation learning, but the pipeline is modular and allows replacement with other time-series transformer variants where operational constraints or pretraining resources differ.

2.2.2 Transformer Applications in Spatio-Temporal Domains

The use of transformer-based models in spatio-temporal analytics extends beyond maritime tracking. In aviation, transformers have been employed to predict aircraft trajectories based on ADS-B feeds, capturing nonlinear motion influenced by air traffic control, weather, and aircraft intent [12]. These models leverage attention mechanisms to weigh the influence of altitude changes or turbulence on trajectory deviations, a technique adaptable to maritime scenarios like storm avoidance. In the Internet of Things (IoT), time-series transformers have enabled real-time segmentation and classification of multi-sensor streams, with successful applications in activity recognition and anomaly detection [13]. In the healthcare domain, transformers have been used to model longitudinal patient data, predicting events such as sepsis or cardiac arrest from vital signs and lab results [14].

These cross-domain successes highlight the generality and flexibility of transformer architectures for time-series modelling, supporting the approach taken in this thesis. Notably, the maritime domain shares similarities with healthcare in its emphasis on rare, high-stakes events (e.g., piracy vs. sepsis), necessitating architectures that balance sensitivity to outliers with robustness to noise.

2.3 Clustering for Behavioural Pattern Discovery

2.3.1 Unsupervised Learning and Clustering Algorithms

Clustering is a cornerstone of unsupervised learning, enabling the discovery of latent patterns without the need for labelled data. In the context of trajectory analytics, clustering is used to group temporally or behaviourally similar segments for higher-level semantic interpretation. Traditional approaches include distance-based methods such as K-Means and DBSCAN, which have been adapted to work on engineered features or trajectory similarity matrices.

However, recent work has shown that clustering over learned embeddings can improve behavioural separation and generalization. Embeddings produced by transformer encoders are particularly suitable for this task due to their capacity to capture high-level semantic patterns across multiple timesteps. In the proposed framework, KMeansConstrained [15] is used to impose minimum cluster size constraints. This not only avoids trivial clusters but ensures a degree of balance that is beneficial for downstream Active Learning and classification.

Density-based methods such as DBSCAN and its hierarchical variants can be attractive in trajectory analytics because they adapt to local density and can discover arbitrarily shaped clusters. Their strengths make them useful in coastal or inland waterways where behaviour densities vary sharply. Nevertheless, these methods have practical drawbacks for large-scale AIS deployments: their performance is sensitive to parameter choices (for example epsilon), and computational complexity grows unfavourably when applied to millions of points. For these reasons, density-based algorithms were evaluated during method development but found less suitable for the combined constraints of scale, the need for reproducible cluster sizes for AL, and heterogeneous sampling rates. Hybrid strategies are possible, for example using density-based clustering to post-process or refine k-means clusters in regions of interest, and such approaches remain promising for targeted operational tasks.

2.3.2 Behavioural Interpretation and Cluster Validation

An essential component of unsupervised analysis is validating and interpreting the discovered clusters. This is especially important in operational domains such as maritime surveillance, where interpretability directly impacts decision-making. In this context, clusters must align with known operational categories, such as transit, loitering, anchoring, or port manoeuvring, if they are to be of practical value.

Several methods are used to validate clustering quality. Silhouette scores [16] assess cluster cohesion and separation, while geospatial overlays help confirm behavioural consistency through spatial localization. In addition, external clustering indices provide complementary perspectives: for example Adjusted Rand Index or Normalized Mutual Information can quantify agreement with sparse ground truth when it exists. Downstream classifier-based validation is also informative: training a supervised model on cluster-derived labels and measuring F1, precision and recall on hold-out data tests whether clusters carry discriminative signal useful for prediction. For example, clusters associated with high SOG and linear trajectories may align with open-sea transits, while low-speed, erratic patterns near port areas may indicate docking or anchoring behaviours. In this thesis, we combine these validation layers to demonstrate that clusters are both geometrically coherent and operationally meaningful.

In addition, the use of bag-of-words (BoW) encoding for summarizing behavioural transitions across a track allows for higher-order aggregation and interpretability. By representing a vessel's journey as a histogram of cluster labels, BoW embeddings enable operators to quickly assess behavioural trends without inspecting individual segments. These approaches, when used in combination, facilitate validation of both local (chunk-level) and global (track-level) behavioural representations.

Figure 2.3 illustrates typical clustering outputs using UMAP projection, where clusters correspond to discrete operational modes (e.g., transit vs. loitering).

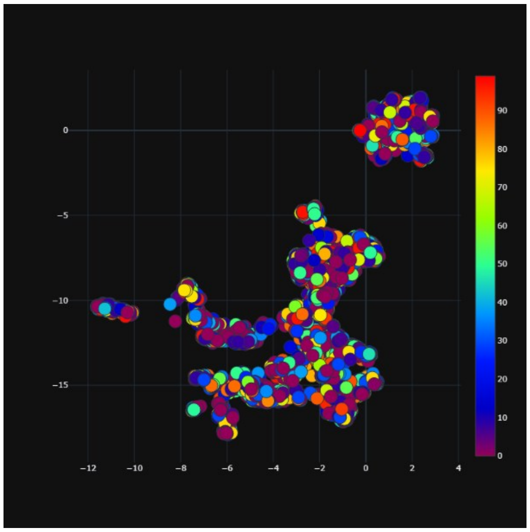


Figure 2.3: UMAP visualization of clustered AIS embeddings with dominant behaviour classes

2.4 Scalable Embedding Storage with Vector Databases

As machine learning systems increasingly rely on learned embeddings for indexing and retrieval, traditional relational databases have proven inadequate due to their inability to support fast, high-dimensional similarity search. Vector databases such as FAISS [17], Milvus [18], and Elasticsearch have emerged to fill this gap.

Elasticsearch, used in both of the systems presented in this thesis [6][7], supports k-nearest neighbour (kNN) search using hierarchical navigable small world (HNSW) indexing. Its hybrid architecture allows it to store metadata alongside vectors and supports filtering, boosting, and multifield querying. This capability enables a variety of analytical tasks: finding similar behaviours across vessels, querying representative samples for annotation, and visualizing embedding clusters.

Vector databases also ease multi-modal integration because they allow storage and filtering of heterogeneous attributes in the same index. In practice, motion embeddings can be stored alongside environmental metadata (for example weather, sea state), port attributes, or derived event features from proprietary engines. Multi-modal fusion can then be realised either by concatenating modality-specific embeddings or by applying cross-attention fusion models; recent multimodal work suggests that cross-modal attention often yields richer joint representations and better downstream performance [24]. The modularity of the vector store therefore supports future extensions to combine AIS with radar, satellite imagery, or environmental sensors.

2.5 Active Learning for Annotation Efficiency

2.5.1 Core Principles of Active Learning

Active Learning aims to reduce labelling costs by selecting the most informative unlabelled samples for annotation. It is especially useful when labelled data is expensive to obtain, and when the marginal value of each new label is significant [19]. In this thesis, four AL strategies are evaluated:

- **Random Sampling:** Baseline approach by uniformly selecting instances.
- **Uncertainty Sampling:** Selecting instances with maximum entropy or minimum prediction confidence.
- **Diversity Sampling:** Maximizing coverage of underrepresented areas in the feature space.

- **Hybrid Sampling:** Combining both uncertainty and diversity to balance exploration and exploitation [20].

These strategies are implemented iteratively to simulate an expert-in-the-loop labelling workflow, retraining models with each iteration and selecting new samples based on the chosen criterion.

2.5.2 Active Learning in Time-Series Contexts

Active Learning in time-series domains poses unique challenges. Unlike static data (e.g., images or text), time-series data includes dependencies across time, variable sequence lengths, and overlapping class boundaries. This complicates uncertainty estimation and diversity quantification. Moreover, time-series classification tasks often require temporal context or aggregation, such as track-level labels derived from segment-level behaviours.

The hybrid strategy used in this thesis addresses these challenges by incorporating both behavioural uncertainty and cluster-aware diversity. Specifically, it selects the most uncertain sample within each coarse-grained cluster, ensuring coverage across the behavioural space while focusing on hard-to-classify examples. This design is particularly effective in the presence of class imbalance because it increases exposure to rare behaviours without sacrificing focus on decision-boundary cases.

Table 2.5 compares the strengths and weaknesses of these strategies in maritime AIS classification.

Table 2.5: Comparison of Active Learning Strategies in Maritime AIS Classification

Strategy	Description	Strengths	Weaknesses
Random	Uniform selection	Simple baseline	Inefficient, redundant sampling
Uncertainty	Highest prediction entropy	Fast convergence	May overfit local uncertainty regions

Diversity	Samples from underrepresented clusters	Ensures coverage	May neglect hard-to-classify samples
Hybrid	Most uncertain sample from each cluster	Balanced, stable performance	Computationally more expensive

2.5.3 AL Performance in Maritime Classification

Empirical validation of the hybrid approach demonstrates clear annotation efficiency gains. For instance, a Macro F1-score of 0.75 is achieved 19 iterations earlier than random sampling. Likewise, model confidence improves more rapidly under hybrid selection. These findings corroborate results in prior AL literature [20], which argue for combined sampling criteria in high-dimensional, imbalanced domains.

2.6 Research Gaps and Contributions

Despite recent advancements, key research challenges remain unaddressed in the domain of spatio-temporal learning for maritime behaviour analysis.

First, most Active Learning research has focused on image and text classification. The adaptation of AL strategies to time-series domains presents unique difficulties due to sequence variability, temporal dependencies, and often multi-label outputs. Temporal chunking exacerbates sample heterogeneity, making sample representativeness and class balance especially difficult to maintain. For example, a single track may contain both "transit" and "illegal fishing" segments, requiring AL strategies to account for multi-label ambiguity.

Second, unsupervised embeddings, while increasingly powerful, suffer from interpretability issues. Most learned vector spaces do not yield immediately meaningful representations, hindering both validation and deployment in mission-critical applications. There remains a lack of accessible, structured representations that bridge learned features with operationally interpretable behaviours.

Techniques like BoW aggregation partially address this but require manual mapping of clusters to semantic labels, a process prone to human bias.

Third, current systems rarely integrate with scalable indexing and feedback mechanisms, limiting their applicability to real-time operational contexts. While vector databases are well suited for similarity search, they are seldom integrated with transformer-based models in the maritime domain, leaving a gap in usable infrastructure. Elasticsearch's integration in this thesis demonstrates that such systems can support real-time querying but highlights the need for standardized maritime-specific indexing schemas.

In response to these gaps, this thesis makes the following detailed contributions:

- **Temporal Representation Learning:** The use of transformer-based time-series embeddings enables the modelling of latent behavioural dependencies within vessel trajectories. Unlike recurrent or convolutional approaches, self-attention mechanisms allow for flexible capture of both short- and long-range motion patterns. This is the first known application of Chronos to segmented AIS data for behavioural encoding.

- **Cluster-Guided Representation and Sampling:** A two-stage constrained clustering strategy is introduced to uncover semantically coherent behaviours. These clusters are leveraged not only for behaviour analysis but also to inform Active Learning sampling, ensuring diversity and reducing sampling redundancy.

- **Bag-of-Words Aggregation for Behaviour Summarization:** By transforming segment-level cluster labels into track-level BoW embeddings, the system bridges unsupervised feature learning with interpretable behavioural classification. This approach enables real-time filtering, visualization, and labelling support with minimal expert involvement.

- **Robust, Scalable AL Loop:** A full Active Learning loop is developed, validated over 100 experimental repetitions. Hybrid sampling reaches key performance thresholds (e.g., F1 ≥ 0.75) up to 32% faster than random sampling. This efficiency enables practical deployment in domains where annotation budgets are limited.

- **Scalable and Interpretable Maritime Analytics:** The system is implemented on real-world, unfiltered AIS data containing class imbalance and behavioural ambiguity. This realistic setting demonstrates that the proposed methods are not only theoretically sound but also operationally viable.

Together, these contributions advance the field of spatio-temporal analytics through a modular and extensible framework that supports real-world decision-making under annotation constraints.

Chapter 3

Unsupervised Embedding and Clustering Pipeline

This chapter introduces the end-to-end architecture used to generate, cluster, and analyze temporal embeddings derived from maritime AIS data. The pipeline automates feature extraction, captures latent behavioural structure, and performs unsupervised grouping of vessel trajectories at scale. Key components include AIS data preprocessing and segmentation, temporal embedding generation using transformer-based models, unsupervised clustering, and a multi-level validation suite encompassing cluster purity, visual separability, and geospatial alignment. Each step is optimized for scalability and interpretability in the context of operational maritime analytics.

3.1 Data Preprocessing and Segmentation

3.1.1 AIS Data Attributes

Raw AIS messages include a mix of static and dynamic fields such as timestamp, latitude, longitude, speed over ground (SOG), course over ground (COG), heading, rate of turn (ROT), vessel type, and draught. These variables are critical in modelling vessel motion and operational context.

Prior to embedding, the AIS dataset is cleaned to remove incomplete or corrupted entries. Missing numerical values (e.g., speed or heading) are imputed using local mean or spline interpolation. Categorical fields like vessel type are encoded using one-hot or ordinal encoding schemes. All features are normalized to ensure scale invariance across dimensions, as recommended in [8]. Data cleaning was non-trivial: AIS transmissions can include duplicate records, irregular timestamps, position jumps due to GPS noise, and gaps caused by transmission outages. Each vessel's time-ordered record is therefore validated for continuity, corrected for out-of-sequence entries, and resampled to uniform temporal intervals. These preprocessing steps were necessary to prevent embedding instability and clustering artefacts caused by noise or missingness.

Table 3.1: Selected AIS features used for embedding generation

Feature	Description	Type
Timestamp	Time of message transmission	Temporal
Latitude, Longitude	Vessel geolocation	Spatial
SOG, COG, ROT	Motion indicators	Numerical
Heading	Vessel heading	Numerical
Vessel Type	Categorical class (e.g., cargo, tanker, fishing)	Categorical
Draught	Dept of ship below waterline	Numerical

3.1.2 Temporal Chunking

AIS data is segmented into 12-hour, vessel-specific time chunks, each representing a temporally coherent behavioural unit. This granularity captures local navigational behaviours (e.g., port entry, anchoring) while maintaining computational tractability. Each chunk becomes an independent instance for embedding.

The 12-hour segmentation window was selected as a balance between temporal context and behavioural purity. Shorter windows increase the number of segments but fragment coherent operations, while longer windows may mix distinct behaviours and blur transitions. Prior work [6] found 12-hour windows sufficient to preserve operational transitions such as port arrival and departure while keeping transformer memory and training cost manageable. Although alternative durations (for example 6 or 24 hours) were tested, performance gains were marginal relative to the additional computational burden.

This segmentation strategy is aligned with operational maritime rhythms and has been empirically validated in prior work [6]. It also reduces the risk of embedding sequences that contain heterogeneous behaviours, thereby enhancing the interpretability of downstream clustering.

3.2 Temporal Embedding Generation

3.2.1 Transformer Architecture

The transformer model employed is based on Chronos [11], a time-series adaptation of the original Transformer encoder [21]. Chronos extends the attention-based design of BERT-style models to multivariate temporal data by applying positional encodings that preserve sequence order and self-attention layers that learn dependencies among features and across time. Each 12-hour AIS sequence is fed as a multivariate time-series into the model, which uses self-attention to learn relationships across timesteps and features.

The input tensors are shaped as $(\boldsymbol{T}, \boldsymbol{F})$ where $\boldsymbol{T}$ is the number of observations in the 12-hour window, and $\boldsymbol{F}$ is the number of encoded features. Positional encodings are added to maintain temporal order. The output is a set of contextualized embeddings, which are then averaged (via mean pooling) to obtain a single fixed-size vector per AIS chunk.

Chronos was selected over other time-series transformers such as Google's TimesFM because of its demonstrated performance on continuous, multivariate sensor data, its open availability, and its relative computational efficiency in handling irregularly sampled series. Moreover, Chronos supports masked forecasting and representation learning within the same framework, which proved beneficial for self-supervised pre-training on AIS sequences.

The final embeddings capture both short-term and long-term behavioural cues such as:

- Repetitive patterns (e.g., loitering near port)
- Transitional dynamics (e.g., acceleration toward open sea)
- Irregularities (e.g., sharp turns, stop-start movements)

Compared to handcrafted features, these representations offer significantly improved performance in clustering, classification, and generalization across domains [6].

3.3 Clustering Methodology

3.3.1 Clustering with KMeansConstrained

Once embeddings are generated, they are grouped using KMeansConstrained [15], a variation of K-Means that enforces a minimum cluster size constraint. This prevents the formation of singleton or overly sparse clusters, which could destabilize subsequent learning steps, especially in Active Learning scenarios.

Clustering is performed in two stages:

1. **Fine-Grained Clustering (k = 100):** This stage produces detailed behavioural groups used for exploratory analysis and chunk-level disaggregation.
2. **Coarse-Grained Aggregation (k = 20):** Related clusters are merged to form broader, interpretable behavioural classes for downstream applications.

The elbow method and silhouette analysis are used to guide the selection of ***k***, while UMAP and t-SNE visualizations confirm separability and consistency within the embedding space. Density-based and probabilistic clustering alternatives, including DBSCAN and Gaussian Mixture Models, were examined during design. However, they exhibited sensitivity to parameter choice and high computational cost at global AIS scale. KMeansConstrained offered a better trade-off between scalability, stability, and interpretability while still supporting later probabilistic analysis if required.

3.4 Behavioural Representation via Bag-of-Words Embeddings

Following the two-stage clustering process, each chunk-level embedding, generated from a 12-hour AIS segment, is assigned a cluster label. These cluster IDs function as behavioural tokens representing localized vessel activity types such as loitering, coastal transit, or anchoring.

To construct a track-level representation, all chunks associated with the same track are grouped, and their corresponding cluster labels are treated as tokens in a behavioural "sentence". This behavioural sentence reflects the vessel's operational progression over time. To encode these sequences into fixed-size, semantically meaningful representations, a bag-of-words (BoW) embedding is computed. This BoW representation quantifies the frequency of each cluster label within a trajectory, producing a sparse, high-dimensional vector that captures the distribution of behavioural states across the entire track.

Formally, if $\boldsymbol{C}$ is the total number of coarse-grained clusters (e.g., 20), then the BoW vector for a given track is:

$$\boldsymbol{b_{track}} = [\boldsymbol{f_1}, \boldsymbol{f_2}, \dots, \boldsymbol{f_C}]$$

where $\boldsymbol{f_i}$ is the frequency cluster $\boldsymbol{i}$ in the sentence formed by the track's chunk labels.

These bag-of-words embeddings are then normalized and used for downstream tasks such as classification, visualization, or sampling in the Active Learning loop (Chapter 4). Unlike raw Chronos embeddings, which represent short-term behaviour at the chunk level, BoW embeddings serve as a compressed summary of vessel behaviour over the full trajectory, making them suitable for higher-level analytics.

This approach bridges the granularity of temporal embeddings with the interpretability of behavioural summarization. It enables both track-level analysis and behavioural clustering, allowing the system to infer high-level patterns across thousands of vessels with minimal manual supervision. It also helps mitigate class imbalance by aggregating rare behaviours at the trajectory level, preventing underrepresentation of minority modes in later supervised or semi-supervised stages.

3.5 System Architecture and Scalability

The entire embedding and clustering pipeline is implemented in a modular, scalable architecture designed to support large-scale AIS analytics. Each system module is detailed below:

3.5.1 Data Ingestion Module

This module handles raw AIS ingestion, parsing, and cleaning. It verifies data consistency and handles errors such as out-of-order timestamps or missing vessel identifiers. After validation, it segments the time-ordered streams into non-overlapping 12-hour windows. Preprocessed sequences are stored in a compressed tabular format compatible with transformer input (e.g., NumPy or TensorFlow datasets).

3.5.2 Embedding Engine

The embedding engine applies the time-series transformer model to each AIS chunk using GPU-accelerated inference. Batched inputs are normalized and embedded in parallel. The model generates 1,000-dimensional output vectors for each chunk, capturing complex motion semantics. Intermediate activations and attention weights are optionally logged for interpretability and debugging.

3.5.3 Vector Database

The vector database, implemented using Elasticsearch, stores all generated embeddings along with their metadata (e.g., timestamp, vessel ID, cluster ID). The system supports efficient k-nearest neighbour search, enabling real-time retrieval of behaviourally similar instances. This functionality is critical for interactive exploration, outlier detection, and diversity-driven sampling in Active Learning.

3.5.4 Clustering Module

The clustering module orchestrates both fine- and coarse-grained unsupervised grouping using K-Means. After clustering, the module logs silhouette scores, cluster size distributions, and intra-cluster variance. Cluster assignments are exported as JSON labels, enabling downstream mapping, annotation, and sampling workflows.

3.5.5 Visualization Interface

A lightweight web-based dashboard enables interactive analysis. The interface supports:

- UMAP projection of embedding space
- Geospatial overlays of cluster centroids
- Per-cluster behavioural summaries
- Similarity search and time-series inspection

This visual tooling supports both automated pipeline validation and human-in-the-loop analysis by domain experts.

Parallel computation is leveraged across modules using multiprocessing and GPU scheduling. Batch caching ensures low-latency reprocessing for iterative experimentation. Overall, the system can process tens of thousands of AIS segments in under ten minutes on a single-node compute server. Figure 3.5 illustrates the system architecture, including data flow and module interdependencies.

While this implementation focuses on AIS-only inputs, the modular architecture readily extends to multi-modal data fusion. For instance, environmental or radar data can be integrated by embedding each modality separately and storing them within a shared vector index for cross-modal similarity search, a common approach in multimodal retrieval systems [24].

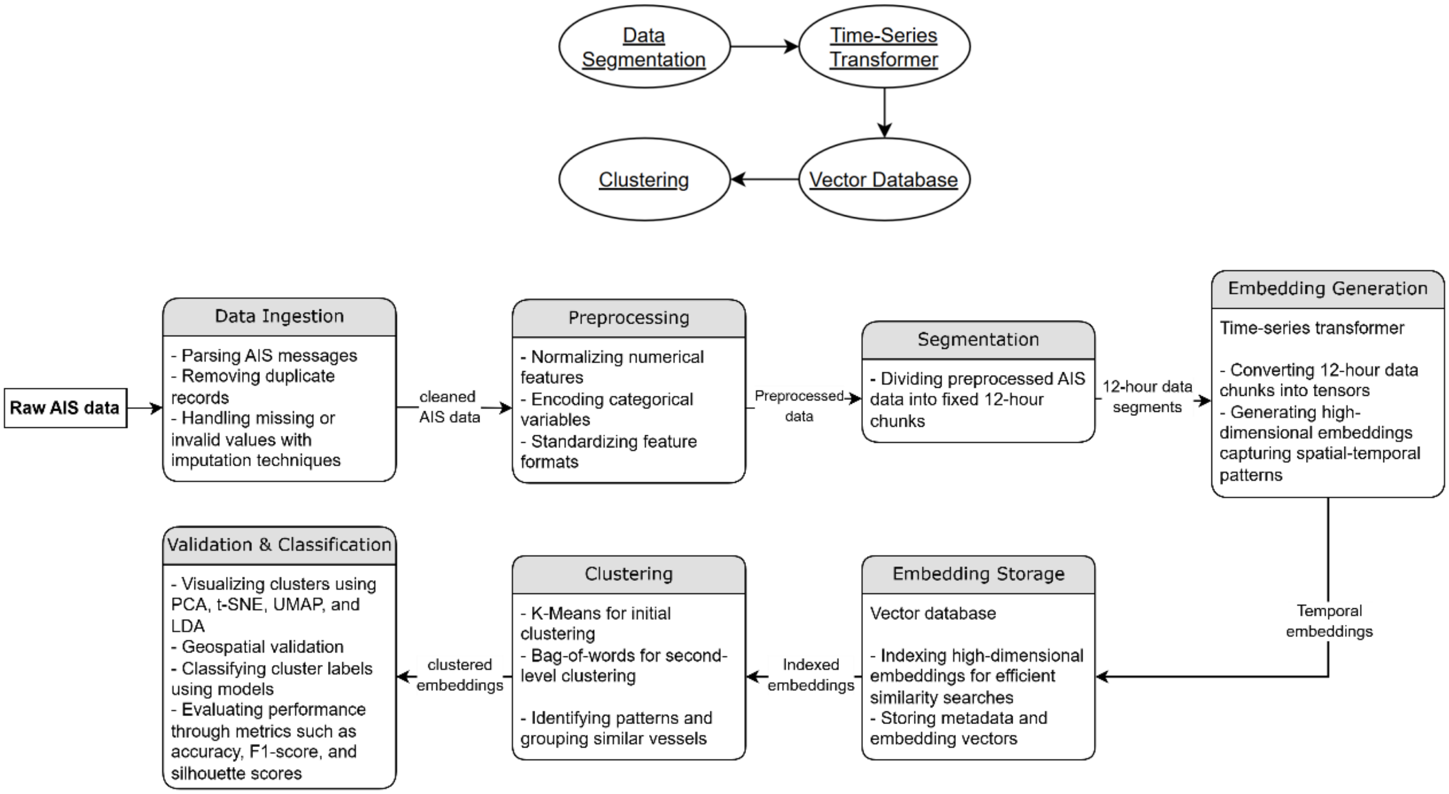


Figure 3.5: System architecture diagram for unsupervised embedding and clustering pipeline

3.6 Validation and Evaluation

3.6.1 Clustering Metrics

Three complementary metrics are used to evaluate clustering quality:

- **Silhouette Score:** Measures intra-cluster cohesion vs. inter-cluster separation [16].
- **Inertia:** Quantifies the sum of squared distances within clusters.
- **Purity:** Measures alignment of cluster labels with human-annotated ground truth (where available).

Results indicate that the 100-cluster solution yields better cohesion and separation (mean silhouette = 0.363), while the 20-cluster setting improves interpretability at the cost of reduced separation (mean silhouette = 0.076). These measures confirm that embedding-based clusters exhibit meaningful structure rather than random partitioning.

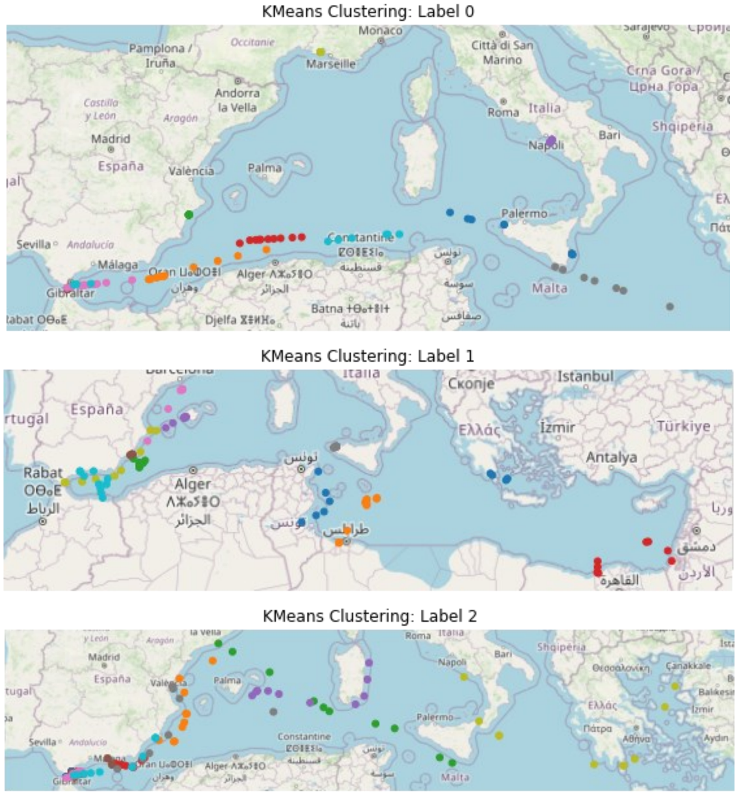
Table 3.6: Clustering validation metrics

Clusters (k)	Silhouette Score	Inertia	Avg. Purity
100	0.363	Low	High
20	0.076	Medium	Moderate

3.6.2 Geospatial Alignment

To validate behavioural realism, cluster centroids are visualized on maritime charts. Spatial patterns associated with each cluster are examined, clusters often correspond to well-defined zones such as coastal corridors, port entry regions, or riverine routes.

Figure 3.6 shows representative overlays from clusters mapped to open sea, river traffic, near port, and along coast behaviours. This geospatial validation is complemented by quantitative checks. For example, Random Forest classifiers trained on cluster-derived labels achieve weighted F1-scores above 0.9 [6], confirming that clusters reflect discriminative behavioural features beyond geographic coincidence. Such internal and external validations jointly demonstrate that the embedding space encodes both motion and contextual semantics effectively.


KMeans Clustering: Label 0

KMeans Clustering: Label 1

KMeans Clustering: Label 2

Figure 3.6: Geospatial plots of representative clusters

3.7 Summary

This chapter detailed the unsupervised pipeline developed to extract temporal embeddings from AIS data and cluster them into semantically meaningful behavioural groups. Leveraging transformer-based models for feature learning, constrained clustering for balanced pattern discovery, and robust validation techniques, the framework delivers scalable and interpretable maritime analytics.

The learned embeddings form the basis for subsequent components in this thesis, including the integration of Active Learning strategies and iterative classification with expert-in-the-loop feedback. The next chapter extends this unsupervised foundation into an interactive annotation pipeline that optimizes label efficiency and model performance.

Chapter 4

Active Learning Framework for Annotation-Efficient Maritime Classification

This chapter introduces an Active Learning extension to the previously described unsupervised embedding and clustering pipeline. The objective of the AL framework is to reduce the cost of manual annotation in large-scale, multi-class classification tasks by intelligently selecting the most informative samples for labelling. Leveraging the bag-of-words (BoW) track representations introduced in Chapter 3, this framework implements an iterative learning loop that maximizes classification performance while minimizing the number of labelled samples required.

The AL system is built upon three core components: transformer-derived embeddings that capture spatio-temporal dynamics, cluster-informed diversity sampling to ensure representational balance, and model-driven uncertainty metrics to guide efficient exploitation. This chapter details the dataset configuration, sampling strategies, classifier design, and the iterative training pipeline. Additionally, it situates the approach within broader AL research, emphasizing innovations tailored to time-series and maritime data.

4.1 Motivation and Problem Setup

Manual annotation of spatio-temporal vessel behaviour is both labour-intensive and highly domain-specific. Despite the utility of high-resolution AIS data, behavioural categorization (e.g., near-port anchoring, riverine navigation) requires contextual understanding of marine operations. The cost and expertise involved in generating ground-truth labels limit the scalability of supervised learning models in operational settings.

Active Learning offers a principled solution by identifying and querying only the most valuable unlabelled samples [19]. Through iterative feedback, it builds performant models from small, strategically chosen subsets of data. This paradigm is especially relevant in maritime analytics, where rare behavioural classes (e.g., inland navigation or anomaly events) are underrepresented and costly to annotate.

In this framework, each AIS track is embedded using the Chronos transformer model and then abstracted into a BoW vector, as described in Chapter 3. These BoW representations serve as input to the classifier and the AL selection process. By iteratively selecting and annotating tracks, the model is incrementally improved while limiting annotation volume. Chronos, a time-series transformer architecture based on self-attention [11], is designed to model dependencies across time without recurrence. It captures both short-term and long-term vessel dynamics, providing contextual embeddings that are robust to noise and variable sequence length. Chronos was preferred over alternatives such as Google's TimesFM due to its demonstrated efficiency on multivariate continuous data, open model accessibility, and compatibility with the downstream clustering and AL modules developed in this thesis.

4.2 Dataset Configuration and Clustering Structure

The AL framework is evaluated on a real-world AIS dataset of 5,000 tracks. Each track corresponds to a unique vessel trajectory and is segmented into 12-hour intervals. Temporal embeddings for each segment are generated using Chronos, and the resulting chunk-level vectors are clustered into 50 behavioural groups using KMeansConstrained (Chapter 3), capturing a diverse array of operational patterns.

In alignment with best practices in AL literature [19], the dataset is split as follows:

- **Initial Labelled Set:** 500 tracks (10%), used to bootstrap the first model. This 10% benchmark is a common heuristic in AL studies, ensuring a representative initial sample while preserving the majority of data for querying.

- **Unlabelled Pool for AL Sampling:** 3,500 tracks, source of samples for AL selection over 70 iterations.

- **Fixed Test Set:** 1,000 tracks (20%), held constant for evaluation throughout the process.

Importantly, the labelled and test sets are stratified across behavioural clusters to ensure balanced representation and to reduce the impact of class imbalance during evaluation. Although real-world maritime data naturally exhibits imbalance (for example transit behaviours vastly outnumber rare port manoeuvres), stratified sampling mitigates early bias and supports more stable convergence across AL strategies.

The 12-hour segmentation window, carried over from the unsupervised pipeline, offers an effective compromise between behavioural resolution and computational feasibility. Shorter windows (for example 6 hours) yield more granular temporal patterns but fragment coherent operations, while longer windows reduce the number of training instances and risk blending heterogeneous activity. This trade-off was evaluated empirically, and no statistically significant gain was observed beyond 12-hour segments.

4.3 Bag-of-Words Embedding Representation

Each vessel track is represented by a BoW embedding derived from the sequence of cluster labels assigned to its constituent 12-hour segments. These BoW vectors act as compressed behavioural summaries, capturing the frequency distribution of cluster-level events over the track's duration.

This abstraction from high-dimensional Chronos embeddings to BoW vectors enables interpretable and efficient classification. It also aligns with prior work on document representation in NLP and has shown success in encoding time-series behaviours into sparse, informative formats [6].

BoW embeddings serve a dual purpose: they are used both as features for classification and as the basis for sample selection in AL. Their fixed dimensionality and interpretability make them ideal for entropy-based and diversity-based sampling strategies.

4.4 Active Learning Sampling Strategies

To assess annotation efficiency and model performance under different querying heuristics, four AL sampling strategies are implemented:

4.4.1 Random Sampling (Baseline)

This strategy selects 50 tracks uniformly at random from the unlabelled pool. It does not leverage any knowledge about the feature space or model performance. While simplistic, it serves as a lower-bound baseline in terms of sampling efficiency and provides a control for evaluating the added value of more sophisticated methods.

4.4.2 Uncertainty-Based Sampling

Uncertainty sampling prioritizes samples near the decision boundary of the current model. Specifically, the entropy of the softmax output is used to quantify prediction uncertainty:

$$\boldsymbol{H(x) = -\sum_{i=1}^{C} P(y_i|x) \log P\,(y_i|x)}$$

where $\boldsymbol{P(y_i|x)}$ is the predicted probability of class $\boldsymbol{i}$ for sample $\boldsymbol{x}$, and $\boldsymbol{C}$ is the number of classes.

This approach tends to improve calibration and model robustness, especially in early AL iterations [19], but may focus disproportionately on ambiguous samples from common classes.

4.4.3 Diversity-Based Sampling

To counteract overfitting and improve generalization, diversity sampling draws samples uniformly across the 50 coarse-grained behavioural clusters. This strategy is designed to ensure representative coverage of the input space and expose the model to a broad range of behaviours.

By incorporating structural diversity from unsupervised clustering, the method mitigates the risk of sampling redundancy and ensures minority behaviours are included.

4.4.4 Hybrid Uncertainty-Diversity Sampling

The hybrid strategy merges the strengths of uncertainty and diversity sampling. In each iteration, the most uncertain sample is selected from each cluster. This ensures that exploration (across behavioural modes) and exploitation (near decision boundaries) are jointly optimized.

This strategy, first introduced in [20], has been shown to improve both label efficiency and stability in iterative learning. It is particularly well-suited for datasets with imbalanced class distributions and behavioural heterogeneity, such as AIS data.

The hybrid uncertainty-diversity strategy remains the most effective method in this framework. Probabilistic sampling approaches, such as Bayesian Active Learning by Disagreement (BALD) or Monte Carlo dropout-based uncertainty estimation, were considered as potential extensions. However, their computational cost scales poorly with the number of AL iterations and the size of the BoW feature space. The implemented entropy-based uncertainty metric therefore provides a practical balance between theoretical rigor and scalability while maintaining statistical interpretability.

4.5 Model Architecture for Classification

The AL system uses a lightweight feedforward neural network to classify track-level BoW embeddings into five behavioural categories:

- Open Sea
- Near Port
- Along Coast
- River Traffic
- Other

The architecture is intentionally kept lightweight to support fast retraining in each AL iteration. Cross-entropy loss is used for training, and the model is optimized using the Adam algorithm with decaying learning rate. Each AL iteration retrains the model from scratch using the current labelled set. This prevents bias accumulation and allows fair comparison of learning efficiency across sampling methods.

Although this thesis does not assign explicit monetary cost to misclassification, its implications are contextually significant. In operational monitoring, false negatives (for example failing to detect anomalous or near-port behaviour) could translate into missed safety alerts or regulatory oversights, while false positives increase manual inspection burden. Hence, the use of Macro F1 and Macro Precision as evaluation metrics ensures sensitivity to both rare and critical behaviours, aligning the model’s performance with real-world operational priorities.

Table 4.5: Neural model architecture

Layer	Type	Input Shape	Output Shape	Activation
Input	Input	100	100	—
Hidden Layer 1	Dense	100	64	ReLU
Hidden Layer 2	Dense	64	32	ReLU
Output	Dense	32	5	Sigmoid

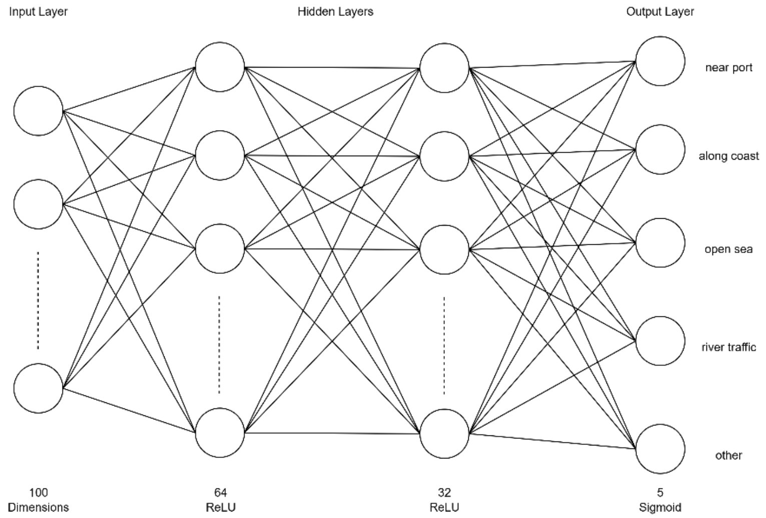


Figure 4.5: Sequential model architecture with an input layer, 2 hidden layers, and an output layer

4.6 Iterative Learning Process

The full AL loop is executed over 70 iterations, querying 50 new samples per round, for a total of 3,500 labelled tracks. Each iteration follows a four-step process:

1. **Model Training:** The classifier is trained on the current labelled set.
2. **Sample Querying:** A batch of 50 unlabelled samples is selected using the active strategy.
3. **Simulated Annotation:** Ground-truth labels for queried samples are revealed (automated in this study).
4. **Evaluation:** The model is evaluated on the 1,000-track test set using predefined metrics.

This loop is repeated across 100 independent trials to account for random seed effects and provide statistically robust results. Aggregated performance metrics and standard deviation bounds are used to compare methods and are qualitatively validated by inspecting the behavioural diversity of queried samples across iterations, confirming that the AL process effectively spans both common and rare operational contexts.

Figure 4.6 illustrates the AL system pipeline, showing the flow between embedding, querying, training, and evaluation modules.

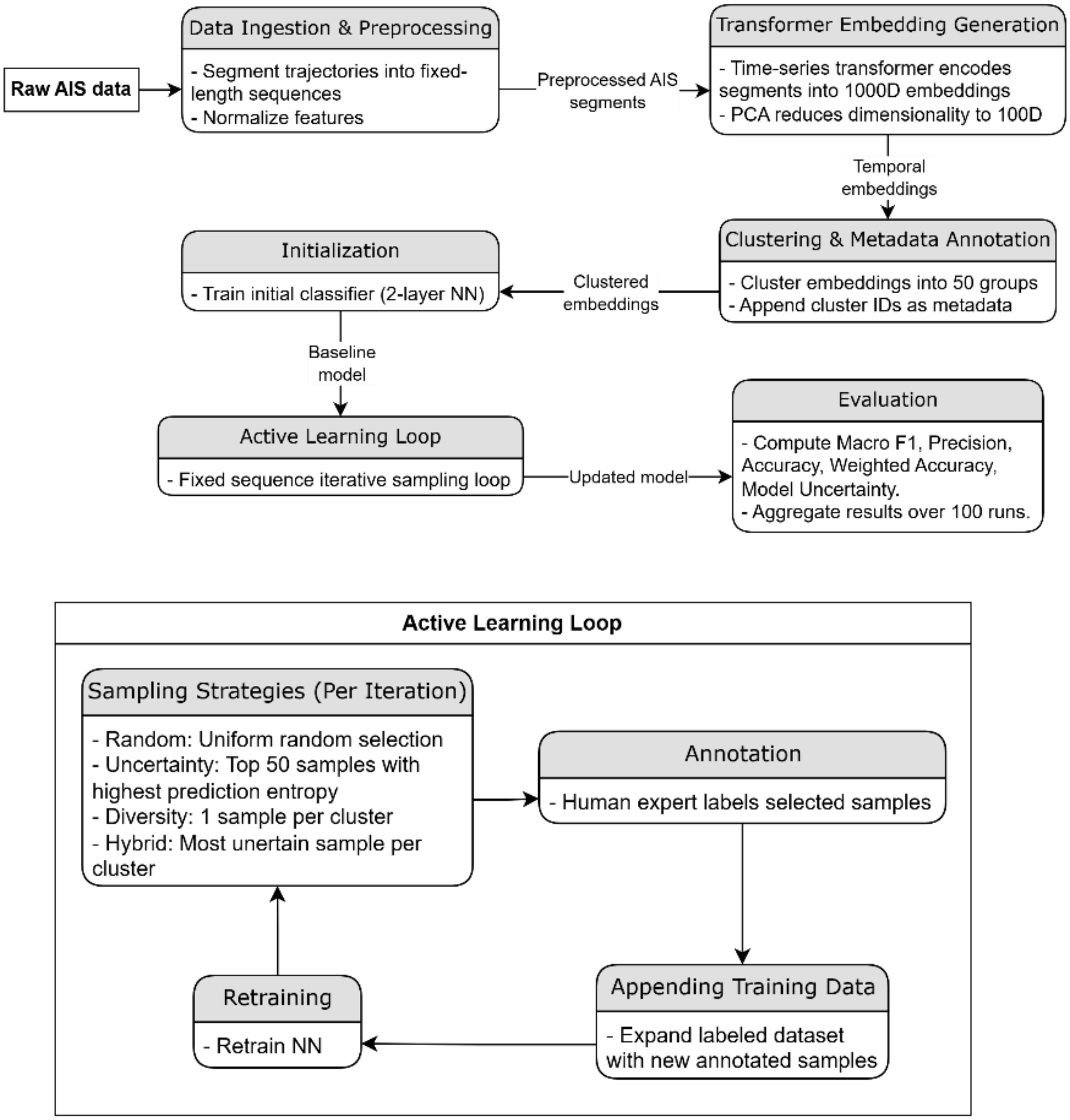


Figure 4.6: System architecture showing the key modules, with a breakdown of the AL loop steps for every iteration

4.7 Evaluation Metrics

To assess both predictive accuracy and annotation efficiency, five key metrics are tracked throughout the AL process:

- **Macro F1-score:** Balances precision and recall across all classes, regardless of size.
- **Macro Precision:** Class-level average precision, highlighting performance on minority behaviours.
- **Accuracy:** Percentage of correctly predicted labels.
- **Weighted Accuracy:** Accuracy adjusted by class frequency in the test set.
- **Model Entropy:** Mean entropy of predictions over the test set, used to monitor calibration.

Table 4.7: Definitions of evaluation metrics

Metric	Definition	Purpose
Macro F1-score	Harmonic mean of macro precision & recall	Class balance, robustness
Macro Precision	Mean precision across all classes	Minority class sensitivity
Accuracy	Total correct predictions / total predictions	Overall performance
Weighted Accuracy	Accuracy scaled by class support	Class imbalance compensation
Entropy	Avg. entropy of model predictions	Confidence and calibration tracking

4.8 Summary

This chapter presented the design and implementation of an Active Learning framework for annotation-efficient maritime behaviour classification. Leveraging transformer-derived embeddings and BoW behavioural summaries, the system integrates four sampling strategies to compare annotation effectiveness.

The proposed hybrid strategy, which selects the most uncertain sample from each cluster, provides a principled balance of exploration and exploitation. The system is evaluated on a real-world dataset with realistic class imbalance and trajectory complexity, making it robust and operationally relevant. Its modular design also permits future extensions to multi-modal settings, where environmental or radar data could be jointly embedded and indexed using the same AL-driven querying logic.

The next chapter presents detailed results across 100 experimental runs, evaluating performance convergence, annotation efficiency, and behaviour-wise model robustness.

Chapter 5

Experimental Validation and Results

This chapter presents a comprehensive empirical validation of the two key contributions developed in this thesis: (1) an unsupervised embedding and clustering pipeline based on time-series transformer embeddings (see Chapter 3), and (2) an Active Learning (AL) framework designed to enable annotation-efficient maritime behaviour classification (see Chapter 4).

The experiments were conducted using a real-world AIS dataset consisting of 5,000 vessel tracks. Each track was segmented into 12-hour temporal windows, transformed into embeddings using a time-series transformer, and subsequently analysed through clustering and classification stages.

In addition to standard AIS-derived features (e.g., speed, heading, rate of turn), this study incorporates features derived from the event output Elasticsearch index using *Total::Insight*™'s patented engine [22]. These proprietary features, fused with core AIS variables, produce a richer representation of maritime behaviour, capturing higher-order context such as proximity to restricted zones or typical activity cycles. This augmented feature set enhances the semantic coherence of embeddings and increases sensitivity to operational anomalies, supporting more accurate and interpretable clustering outcomes.

Data cleaning for this dataset was moderately complex due to noise, missing transmissions, and out-of-order timestamps commonly found in raw AIS streams. Standard preprocessing included filtering erroneous coordinates, interpolating missing temporal values, and normalizing motion variables such as speed and rate of turn. These steps were automated to ensure reproducibility while maintaining data fidelity, given that over-filtering could inadvertently remove operational anomalies of interest.

The evaluation is structured into two major sections:

- Section 5.1 evaluates the quality of unsupervised representations and clustering, including classification alignment and geospatial consistency.
- Section 5.2 presents the results of the AL framework across four sampling strategies, focusing on convergence dynamics, label efficiency, and model robustness under real-world class imbalance.

Where relevant, results are contextualised within the broader research landscape and supported with figures, tables, and insights from the accompanying publications.

5.1 Unsupervised Embedding and Clustering Evaluation

5.1.1 Embedding Coherence and Clustering Validity

The temporal embeddings were generated using Chronos, a transformer-based time-series model that learns temporal dependencies through self-attention [11]. Chronos was selected over alternatives such as Google's TimesFM because it is optimized for multivariate sensor data, supports fine-grained contextual encoding, and is compatible with large unlabelled datasets typical of AIS streams. Unlike recurrent models, Chronos processes entire temporal sequences in parallel, effectively capturing both short-term manoeuvres (e.g., port entry) and long-term trends (e.g., open-sea cruising) without requiring handcrafted temporal features.

The embeddings were evaluated using a two-stage clustering strategy with K-Means. Two configurations were assessed:

- A fine-grained setting with $k = 100$, intended to capture detailed behavioural nuances,

- A coarse-grained setting with ***k*** = 20, selected for interpretability in downstream annotation tasks.

The optimal values of ***k*** were determined using the elbow method:

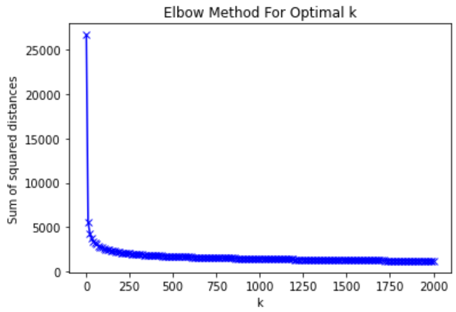


Figure 5.1.1A: Elbow method plot for fine-grained clustering (k = 100)

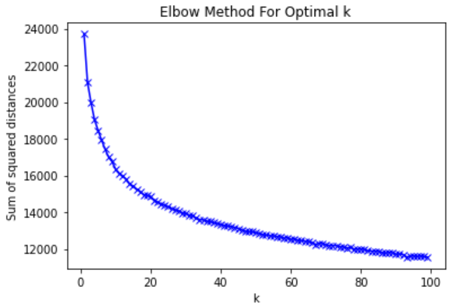


Figure 5.1.1B: Elbow method plot for coarse-grained clustering (k = 20)

Cluster cohesion and separation were further evaluated via silhouette analysis:

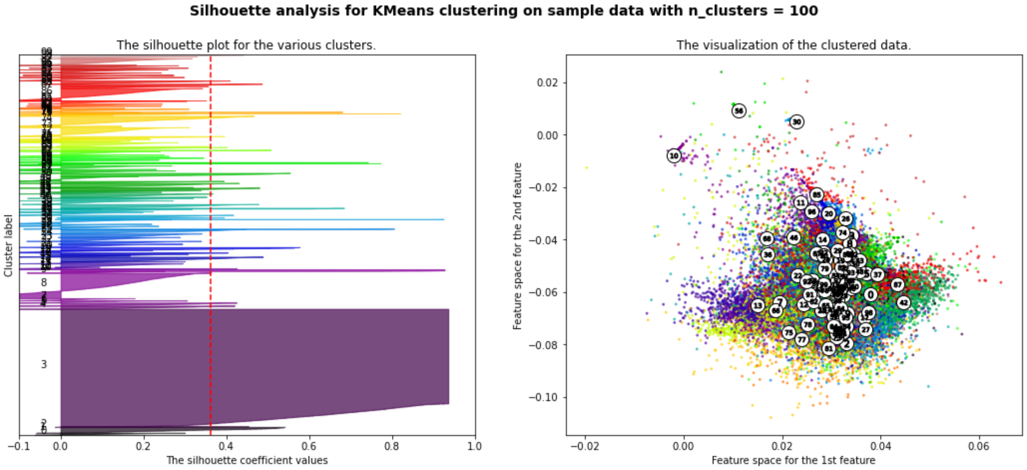


Figure 5.1.1C: Silhouette analysis for KMeans clustering (k = 100)

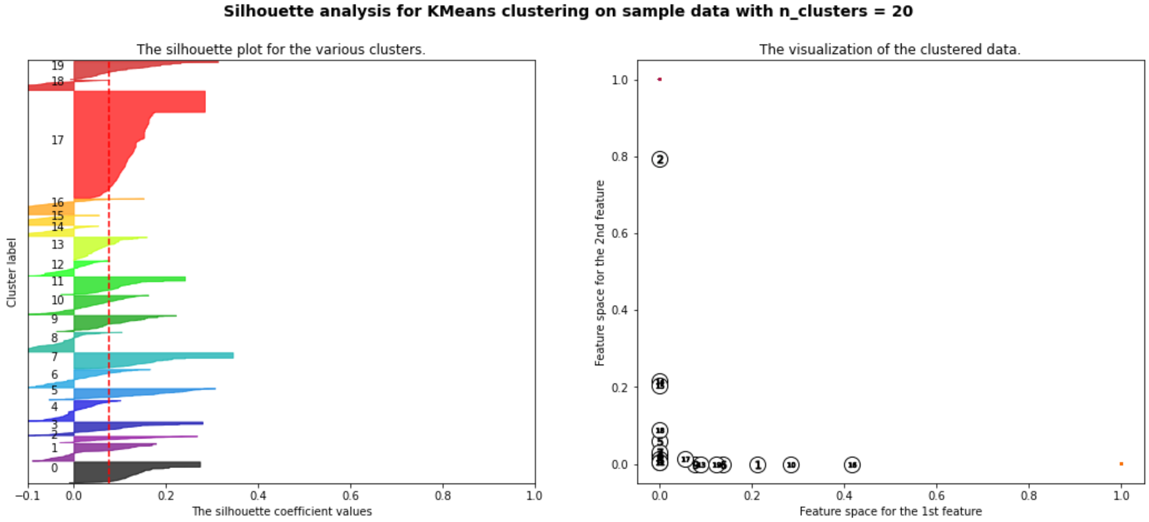


Figure 5.1.1D: Silhouette analysis for KMeans clustering (k = 20)

The mean silhouette score for the fine-grained setting was 0.363, indicative of well-separated clusters with strong intra-cluster similarity. This suggests that the transformer embeddings effectively captured temporal and spatial structure. By contrast, the coarse-grained clustering produced a silhouette score of 0.076, reflecting partial overlap between general behavioural categories, a trade-off deemed acceptable given the improved interpretability and cluster-level behavioural labelling.

Although K-Means was chosen for its scalability and interpretability, other density-based techniques, such as DBSCAN or HDBSCAN, were considered for exploratory validation. These methods produced reasonable behavioural separation but were computationally prohibitive at scale and highly sensitive to local density fluctuations in AIS coverage. Future work could revisit these methods using distributed implementations to explore density-adaptive behaviour grouping.

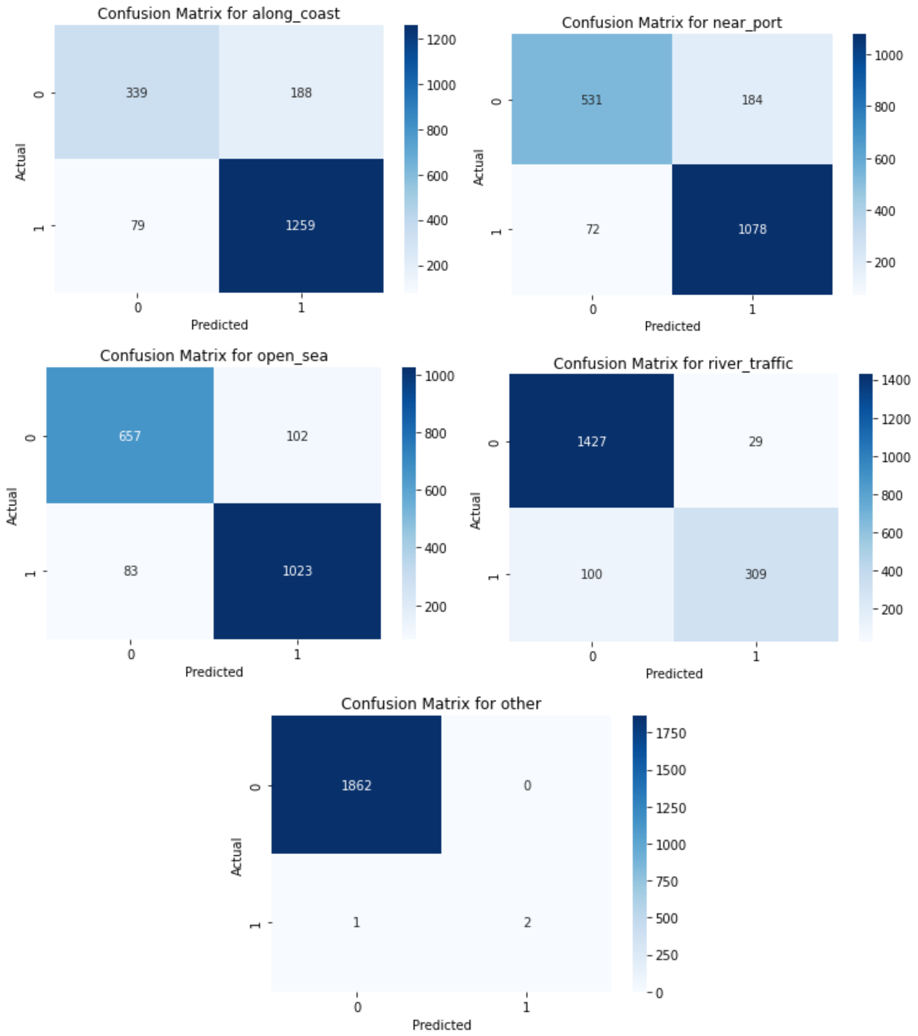
These results are consistent with prior work in unsupervised trajectory analysis [1][2], which observed similar trade-offs between granularity and semantic labelling.

5.1.2 Random Forest Classification and Cluster Annotation Consistency

To evaluate whether unsupervised clusters captured semantically meaningful patterns, a Random Forest classifier was trained on the transformer-derived embeddings, with labels derived from the 20 coarse clusters configuration.

Table 5.1: Random Forest classification report for 5 behavioural classes

Class	Precision	Recall	F1-score	Support
Along Coast	0.87	0.94	0.90	1338
Near Port	0.85	0.94	0.89	1150
Open Sea	0.91	0.92	0.92	1106
River Traffic	0.91	0.76	0.83	409
Other	1.00	0.67	0.80	3
Weighted Avg.	**0.88**	**0.92**	**0.90**	**4006**



Confusion Matrix for along_coast

Confusion Matrix for near_port

Confusion Matrix for open_sea

Confusion Matrix for river_traffic

Confusion Matrix for other

Figure 5.1.2A: Confusion matrix for Random Forest predictions (per class)

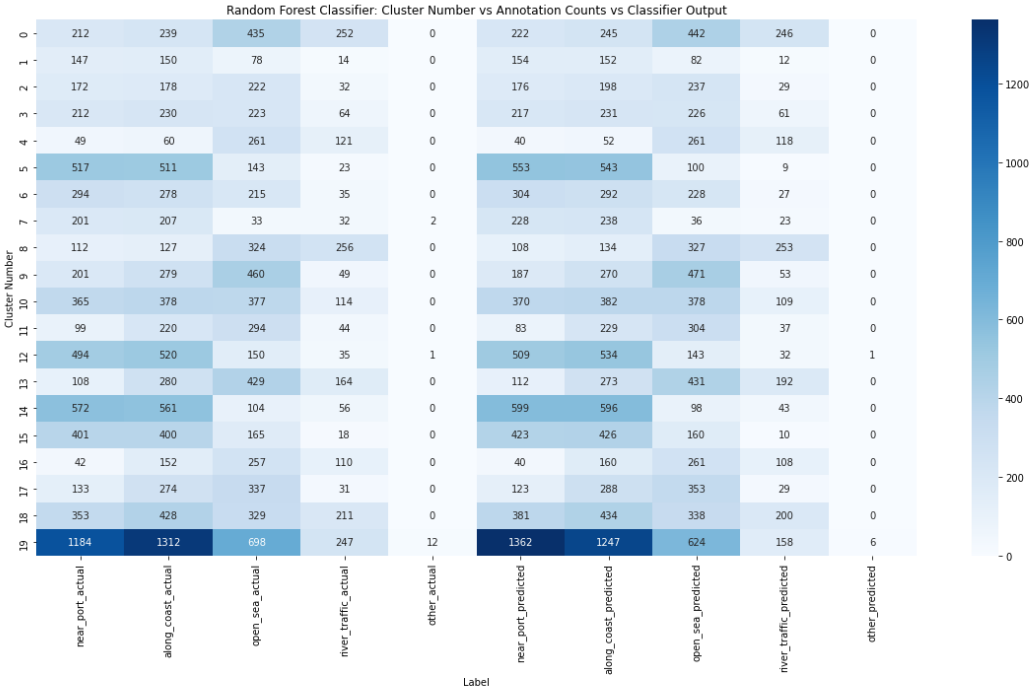
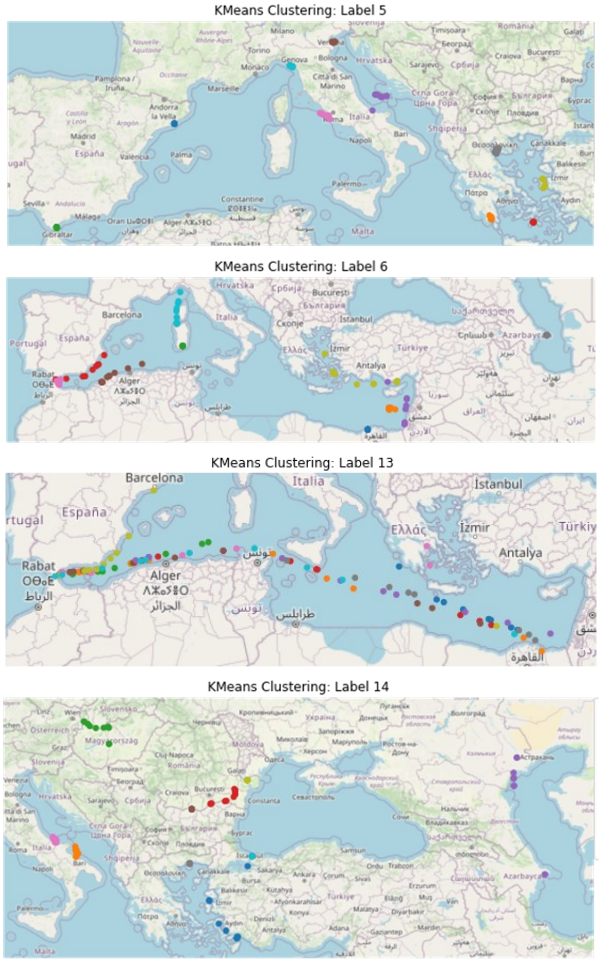


Figure 5.1.2B: Cluster number vs annotation counts vs classifier output (final 20 clusters)

The classifier achieved strong performance across major behavioural categories, with F1-scores ranging from 0.83 to 0.92 for the dominant classes. Despite the extreme rarity of the “other” category (only 3 labelled instances), the classifier generalised moderately well, yielding an F1 of 0.80. This supports the argument that transformer-derived embeddings and cluster-based labelling yield meaningful semantic structure suitable for downstream tasks, validating results reported in [6].

5.1.3 Geospatial Validation

Cluster centroids from the coarse-grained configuration were projected onto a maritime chart, validating alignment with real-world operational zones.



KMeans Clustering: Label 5

KMeans Clustering: Label 6

KMeans Clustering: Label 13

KMeans Clustering: Label 14

Figure 5.1.3: Geospatial overlays of representative clusters, for clusters 5, 6, 13, and 14, showing near port, along coast, open sea, and river traffic respectively

The clusters displayed spatial coherence with known maritime patterns. For example, open sea clusters aligned with known shipping lanes, and river traffic clusters followed inland waterways. This geospatial validation confirms that the unsupervised framework preserves operational semantics and can support real-time decision-making scenarios.

While geospatial overlays confirmed that clusters aligned with known maritime routes and zones, this was complemented by statistical validation and downstream classification performance to ensure that the clusters reflected genuine behavioural differences rather than geographic proximity alone.

5.2 Active Learning Performance Evaluation

5.2.1 Macro Metrics Over Iterations

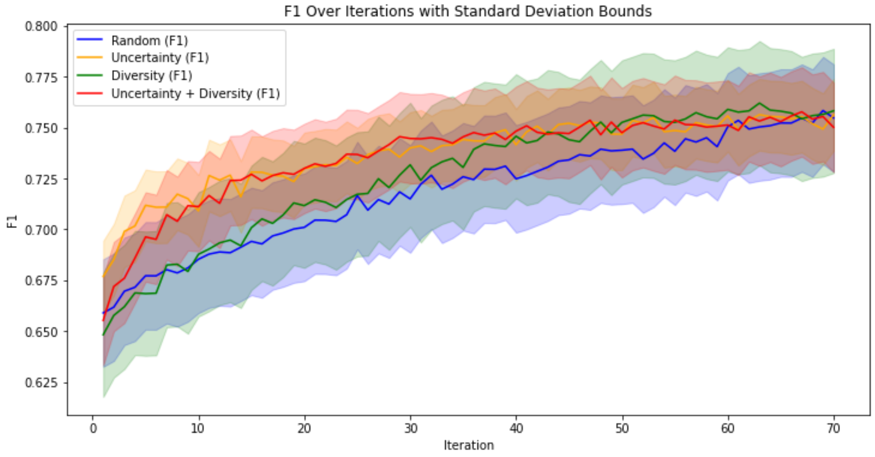


Figure 5.2.1A: Macro F1-score over AL iterations (with standard deviation bounds)

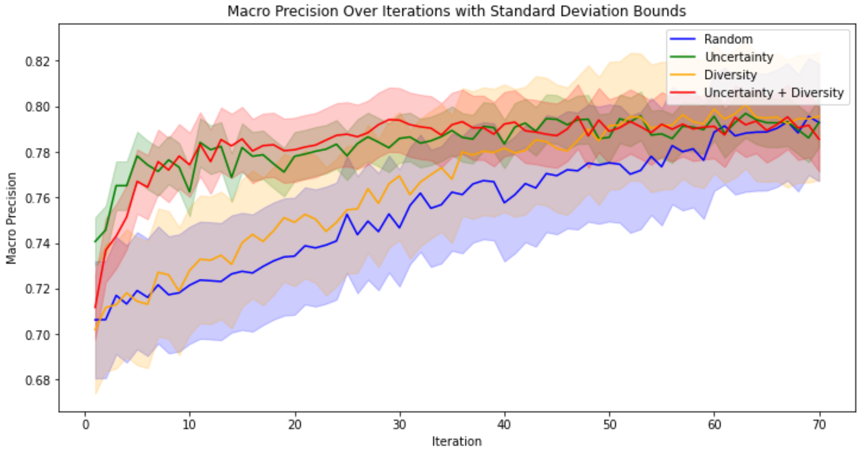


Figure 5.2.1B: Macro Precision over AL iterations (with standard deviation bounds)

Table 5.2.1: Average performance across all 70 iterations (100 runs)

Metric	Random	Uncertainty	Diversity	Hybrid
Macro F1-score	0.7190	**0.7375**	0.7266	0.7360
Macro Precision	0.7554	0.7839	0.7659	**0.7844**
Accuracy	0.8345	0.8341	0.8340	**0.8353**
Weighted Accuracy	0.7813	0.7785	0.7827	**0.7837**
Model Uncertainty (Entropy)	**1.1039**	1.1061	1.1083	1.1172

The hybrid method demonstrated strong average performance across all metrics. While uncertainty sampling slightly outperformed it early in the training process, hybrid sampling achieved competitive accuracy and demonstrated greater consistency across trials. The entropy metric reflects steady reduction in predictive uncertainty, indicative of stable model calibration.

5.2.2 Label Efficiency and Convergence Rates

Table 5.2.2A: Iteration at which key F1 thresholds were reached

Threshold F1	Random	Uncertainty	Diversity	Hybrid
≥ 0.700	19	4	15	7
≥ 0.725	32	11	29	15
≥ 0.750	60	42	48	41

Table 5.2.2B: Iteration at which key macro precision thresholds were reached

Threshold Macro Precision	Random	Uncertainty	Diversity	Hybrid
≥ 0.750	25	3	19	4
≥ 0.775	48	5	36	7
≥ 0.800	61	38	48	28

Across all thresholds, the hybrid method reached high-performance regimes faster than the baseline and diversity-based strategies. For example:

- To reach F1 ≥ 0.75, hybrid required 41 iterations, compared to 60 for random, a 32% reduction in labelled samples.
- To reach macro precision ≥ 0.80, hybrid needed 28 iterations, compared to 61 for random, a 54% reduction in labelling cost.

These improvements reflect findings in recent literature [20], which emphasize that combining behavioural diversity with predictive uncertainty improves label efficiency without sacrificing model generalisation.

5.2.3 Additional Metrics and Model Dynamics

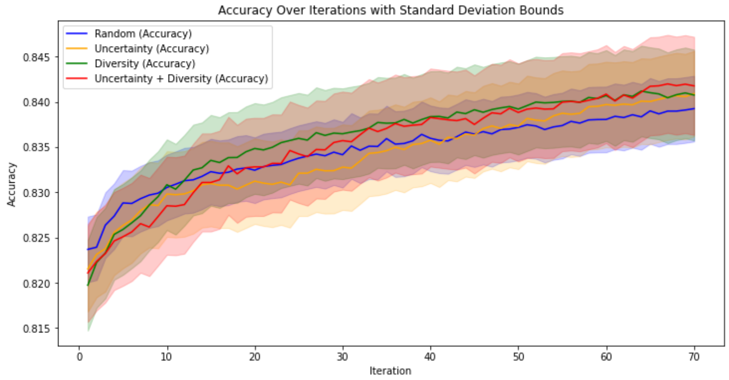


Figure 5.2.3A: Accuracy over AL iterations (with standard deviation bounds)

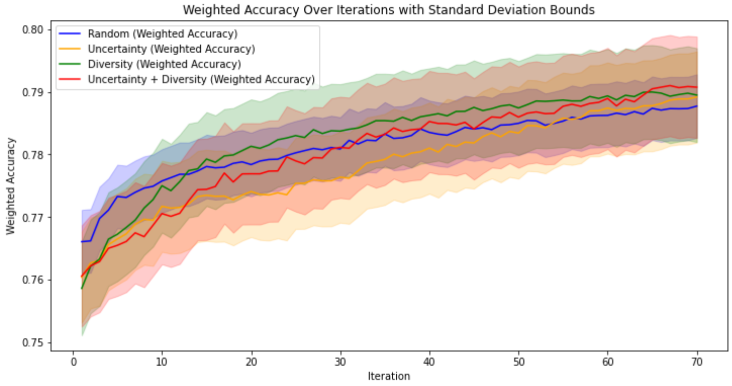


Figure 5.2.3B: Weighted accuracy over AL iterations (with standard deviation bounds)

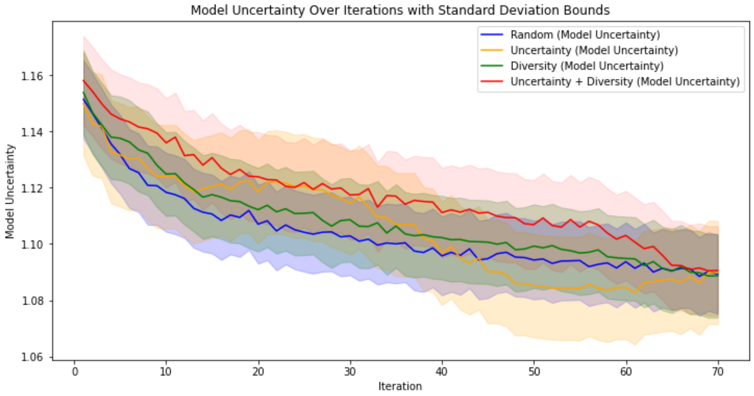


Figure 5.2.3C: Model uncertainty (entropy) over AL iterations (with standard deviation bounds)

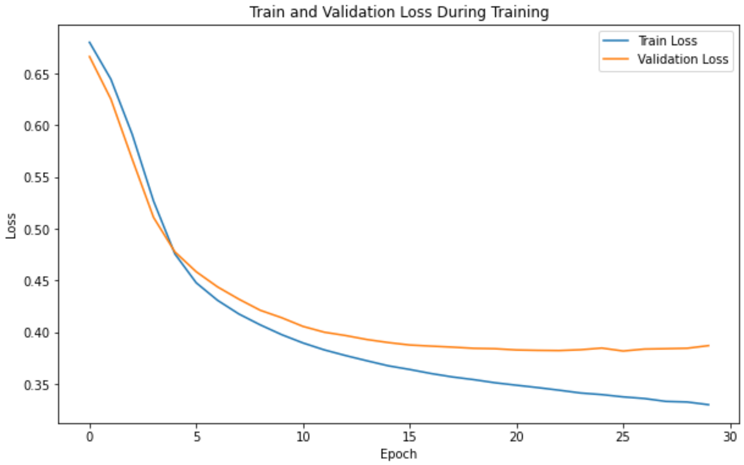


Figure 5.2.3D: Train-test loss curve

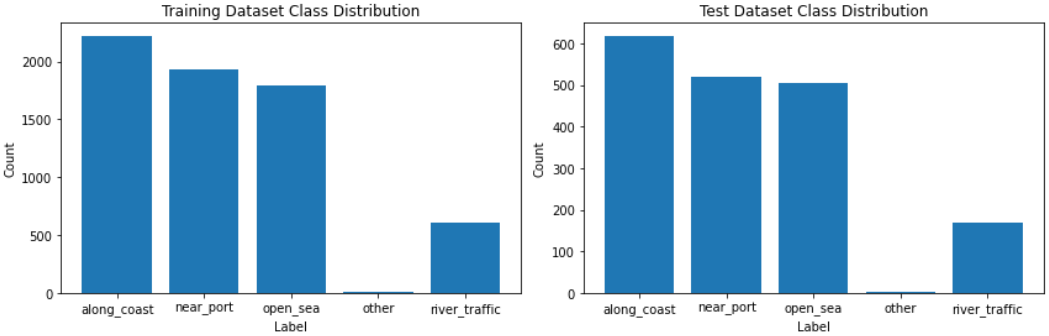


Figure 5.2.3E: Train-test class distribution histogram

The dataset was intentionally left imbalanced to reflect real-world maritime data distributions. For instance, the "other" category appeared only 13 times across 5,000 tracks. Rather than artificially balancing the dataset, this framework demonstrates robustness under these practical constraints. The hybrid strategy was the most robust in handling this imbalance, demonstrating strong F1 scores for rare classes. The consistency of entropy reduction and loss convergence across 100 trials attests to the model's resilience under sparse supervision.

The cost of misclassification varies by context: false negatives in categories like near port or river traffic may correspond to missed detection of safety-critical activity, while false positives can result in unnecessary manual reviews. The AL framework mitigates such operational risks by prioritizing uncertain samples near decision boundaries, where misclassifications are most likely. This strategy optimizes annotation resources for cases that would otherwise yield the highest operational cost if left mislabelled.

5.3 Limitations

5.3.1 Class Imbalance and Minority Representation

While the hybrid AL method improved performance on underrepresented classes, limitations remain when data scarcity becomes extreme. Exploring advanced techniques such as data augmentation, label smoothing, or generative modelling for minority behaviour synthesis would be a logical extension.

5.3.2 Fixed Temporal Windowing

The use of fixed 12-hour segments to produce chunk embeddings simplifies preprocessing and batch processing but may disrupt temporal continuity. Behavioural transitions occurring near window boundaries could result in mischaracterized representations. Adaptive temporal segmentation or hierarchical time-series encoding could provide finer behavioural continuity and improve semantic alignment.

5.3.3 Dependency on Initial Clustering Quality

The quality of bag-of-words representations and diversity sampling relies on the initial clustering. Errors in clustering can propagate, affecting AL performance. Future pipelines may include online re-clustering or embedding space refinement.

5.3.4 Real-World Annotation Considerations

Although ground-truth labels are simulated in this study, real-world annotation is subject to domain expertise variability and fatigue. Integrating interactive labelling tools and semi-supervised validation pipelines may help operationalise these systems effectively.

5.4 Summary

This chapter provided comprehensive empirical validation of two key contributions:

- The unsupervised embedding and clustering framework yields interpretable, coherent groupings of maritime behaviours, validated both geospatially and through downstream classification.

- The Active Learning system achieves state-of-the-art annotation efficiency, with the hybrid sampling strategy outperforming all baselines by significantly reducing labelling costs, while maintaining high classification accuracy and model stability.

Together, these components form a robust foundation for scalable maritime analytics, combining unsupervised structure discovery with interactive learning. The next chapter synthesises the overall contributions, addresses remaining challenges, and proposes avenues for extension across domains and modalities.

Chapter 6

Conclusion and Future Work

This thesis has presented a unified framework for advanced unsupervised and annotation-efficient analysis of spatio-temporal maritime data, addressing core challenges in representation learning, behavioural abstraction, and scalable annotation. By integrating transformer-based temporal embeddings, constrained clustering, and Active Learning, the proposed methodology offers a data-driven alternative to traditional, heuristic-based systems for maritime behaviour analysis.

This concluding chapter synthesizes the key contributions and experimental findings, articulates the broader implications of the work across disciplines, and proposes concrete directions for future research. It aims to position this research within the broader trajectory of spatio-temporal analytics and outline how it can evolve to meet operational demands across domains.

6.1 Summary of Contributions

The thesis advances the state of the art through four interconnected contributions spanning representation learning, clustering-based summarization, efficient labelling, and real-world validation.

6.1.1 Transformer-Based Temporal Embeddings for Maritime Behaviour Modelling

This work proposed a novel use of transformers to encode AIS-based time-series into dense, contextualised embeddings. The Chronos model, a time-series adaptation of the transformer architecture, was selected for its ability to process multivariate sensor data efficiently while maintaining temporal dependencies through self-attention. Chronos generalises well across unlabelled or weakly labelled data streams and was preferred over alternatives such as Google's TimesFM because it is open, lightweight, and optimized for irregularly sampled time-series.

By segmenting AIS tracks into 12-hour windows and transforming them into fixed-size vectors using self-attention, the model captures both short- and long-term behavioural dependencies without relying on domain-specific features. The 12-hour window was empirically chosen as a balance between temporal resolution and computational efficiency. Shorter windows increased noise and fragmented trajectories, while longer windows risked merging distinct behavioural modes.

These embeddings offer key advantages:

- They eliminate the reliance on handcrafted motion descriptors (e.g., SOG, ROT),
- They capture long-range temporal dependencies,
- They are adaptable across vessel types, geographic regions, and operational scales,
- . They support scalable indexing via vector databases, enabling similarity-based querying, clustering, and downstream classification.

By applying this approach to real-world AIS data, the thesis demonstrates that temporal transformers can produce semantically meaningful embeddings for structured but weakly labelled motion data, a finding aligned with emerging work in mobility analytics and environmental sensing [11][21].

6.1.2 Clustering for Unsupervised Behavioural Discovery

The learned embedding space was structured through a two-tiered clustering approach using K-Means, yielding both fine-grained and interpretable behavioural clusters. Validation through silhouette scores, classification metrics, and geospatial overlays confirmed that the clusters preserved semantically coherent behaviours such as near-port manoeuvring, river traffic, and open-sea transit.

While K-Means was chosen for its scalability and interpretability, density-based clustering techniques such as DBSCAN were also considered for exploratory analysis. These alternatives offered useful insight into local trajectory densities but were less stable under large-scale, high-dimensional embeddings, particularly given uneven AIS sampling densities. Future work may revisit such approaches using distributed or adaptive-density implementations.

A key innovation was the transformation of chunk-level cluster assignments into bag-of-words (BoW) vectors summarising behaviour across entire vessel tracks. These vectors enabled robust classification while maintaining interpretability, a critical feature for operational deployment. Similar techniques have seen success in sequential summarisation in surveillance, logistics, and animal telemetry [23].

6.1.3 Active Learning for Label-Efficient Classification

To address the annotation bottleneck, the thesis introduced a transformer-compatible AL framework incorporating four sampling strategies, random, uncertainty, diversity, and a hybrid uncertainty-diversity approach. The hybrid strategy, which selects the most uncertain instance from each behavioural cluster, significantly improved label efficiency.

Key findings include:

- A 32% reduction in labelled samples to reach F1 ≥ 0.75 when compared to random sampling,
- Consistent superiority in macro F1 and macro precision across 100 trials,

- Stability and resilience in the face of real-world class imbalance, including successful classification of rare behaviours like "other".

The experiments intentionally retained data imbalance to reflect real-world maritime conditions. Despite this imbalance, the hybrid strategy achieved reliable convergence and high precision, suggesting strong generalisation and cost-effective annotation. These outcomes reaffirm the growing body of work highlighting AL as a practical tool for scaling supervision in structured and temporally dependent data domains [19].

6.1.4 Operational and Analytical Robustness

In contrast to benchmarked academic datasets, the AIS dataset used here was deliberately unfiltered and imbalanced. This reflects real-world maritime complexity, including missing data, rare behaviour, and heterogeneous operational zones. Data cleaning was therefore non-trivial, requiring the removal of corrupted transmissions, correction of out-of-order timestamps, and imputation of missing motion attributes while retaining operational continuity.

Despite these challenges, the framework remained performant, illustrating that unsupervised and weakly supervised methods, when properly architected, can deliver operationally viable solutions. Validation extended beyond geospatial consistency to include internal metrics and downstream classification performance, confirming that behavioural clusters were meaningful in both spatial and semantic dimensions.

This real-world validation underscores the importance of designing methods that are not only theoretically robust but also deployable under the constraints of messy, sparse, and dynamic environments.

6.2 Broader Implications

The contributions of this work extend beyond the maritime domain. The modular pipeline, comprising temporal transformers, clustering-based summarisation, and Active Learning, is broadly applicable to domains where spatio-temporal structure, limited labelling, and behavioural variability are common.

Potential areas include:

6.2.1 Aviation and Air Traffic Management

Automatic Dependent Surveillance–Broadcast (ADS-B) data for aircraft shares structural similarities with AIS, capturing trajectory dynamics across flight phases. Techniques introduced here could support clustering of aircraft behaviours, such as holding patterns or missed approaches, and AL could be used to selectively annotate complex or novel trajectories in congested airspace [12].

6.2.2 Logistics and Supply Chain Monitoring

Truck, rail, and container movement data share many characteristics with maritime tracks. Behavioural summarization using unsupervised cluster-to-BoW pipelines could support bottleneck detection or delay prediction in multimodal logistics networks. Ground vehicle movement across road, rail, and port networks exhibits temporal dependencies that could be modelled using transformer embeddings. Cluster-informed sampling could support anomaly detection (e.g., unexpected stops or detours) and improve transparency in global supply chains, an area increasingly reliant on automated decision support.

6.2.3 Environmental and Ecological Monitoring

Animal telemetry datasets, particularly those derived from GPS-tagged species, could benefit from this framework. Behaviours such as migration, foraging, or avoidance can be encoded temporally and clustered to reveal patterns linked to seasonal or environmental conditions. Similar approaches have been shown to uncover habitat transitions and threat responses in marine ecology [23].

6.3 Directions for Future Research

Building on the contributions of this thesis, several promising directions emerge for advancing the methodology and expanding its scope.

6.3.1 Multi-Modal Integration

Future work could incorporate additional sensor modalities, such as radar signatures, weather feeds, sonar data, satellite imagery, alongside AIS data. Combining visual, environmental, and motion data may yield richer behavioural models, especially for detecting context-sensitive anomalies.

Embedding architectures such as multi-modal transformers or cross-attention fusion models could enable this integration [24]. This extension would also support multi-modal Active Learning, where information from one modality informs sampling in another, increasing label efficiency across heterogeneous data streams.

6.3.2 Predictive Behaviour Modelling

While this work focuses on retrospective analysis, the embeddings and BoW vectors could support predictive tasks such as route deviation, anchoring likelihood, or compliance breaches. Incorporating sequence models (e.g., attention-based decoders) could transform static summarisation into real-time trajectory forecasting. Future efforts could explore probabilistic transformer variants or Bayesian sequence models to quantify uncertainty in behavioural predictions, thereby supporting decision confidence in safety-critical operations.

These probabilistic extensions would provide not only point predictions but also confidence intervals, directly addressing the cost of misclassification in operational monitoring. For example, false negatives in near-port or riverine categories may conceal collision or regulatory risks, while false positives can generate unnecessary alerts. Probabilistic methods would allow such costs to be integrated explicitly into decision-making workflows. Such extensions would position the framework as a predictive analytics engine capable of powering alerting systems, risk scoring models, and anticipatory operations planning.

6.3.3 Online Learning and Concept Drift

Behavioural norms in maritime traffic shift over time due to regulatory changes, seasonal cycles, and geopolitical factors. Future research could explore streaming Active Learning, online embedding updates, and drift detection to ensure the system remains adaptive. Approaches such as continual learning with rehearsal or memory-based sampling may help preserve historical context while adapting to new behaviours. Implementing incremental retraining of transformer embeddings, paired with streaming Active Learning, could help maintain adaptability without sacrificing performance stability.

6.3.4 Interactive Human-in-the-Loop Interfaces

Deploying the framework in real-world systems will require interpretable, interactive interfaces for analysts. Visual tools that allow experts to explore embeddings, cluster behaviours, and selectively label or correct samples could significantly enhance adoption. This includes:

- Embedding visualizations with geospatial overlays,

- Cluster exploration dashboards,
- Feedback-based refinement of cluster-to-behaviour mappings.

Such interfaces would enhance trust, support decision auditing, and enable collaborative human-AI workflows. Incorporating techniques from explainable AI (XAI) would further support regulatory adoption and user confidence in high-stakes maritime domains. These interfaces could also support probabilistic visualisations of uncertainty to help analysts understand model confidence in real time.

6.4 Concluding Remarks

This thesis has demonstrated that unsupervised and weakly supervised methods, when integrated through thoughtful architecture and informed by domain realities, can rival traditional supervised pipelines in accuracy, interpretability, and efficiency. By coupling transformer-based representation learning with clustering-guided summarization and hybrid Active Learning, this work delivers a robust foundation for real-time maritime behaviour analysis.

Beyond methodological contributions, this research offers a practical blueprint for developing spatio-temporal analytics systems that are interpretable, scalable, and resilient to annotation scarcity. Its deployment on real-world, unfiltered AIS data underscores its operational potential, while its modular design invites adaptation to a wide range of spatio-temporal domains.

Future advances in multi-modal integration, probabilistic modelling, and interactive analytics promise to further enhance both interpretability and predictive power. It is hoped that the contributions made herein will inspire continued innovation at the intersection of Machine Learning, behavioural science, and dynamic system monitoring, particularly in settings where data is abundant, but human labelling remains expensive, inconsistent, or infeasible.

Bibliography

[1] L. van der Maaten, E. O. Postma, and J. van den Herik, “Dimensionality reduction: A comparative review,” Journal of Machine Learning Research, vol. 10, no. 1, 2009.

[2] G. Pallotta, M. Vespe, and K. Bryan, “Vessel pattern knowledge discovery from AIS data: A framework for anomaly detection and route prediction,” Entropy, vol. 15, no. 6, pp. 2218–2245, 2013.

[3] P-R. Lei, “A framework for anomaly detection in maritime trajectory behavior,” Knowledge and Information Systems, vol. 47, no. 1, pp. 189-214, 2016.

[4] H. Duan, F. Ma, L. Miao, and C. Zhang “A semi-supervised deep learning approach for vessel trajectory classification based on AIS data,” Ocean & Coastal Management, vol. 218, 2022.

[5] C. Gormley and Z. Tong, “Elasticsearch: The definitive guide: A distributed real-time search and analytics engine,” O'Reilly Media, Inc., 2015.

[6] P. Saleh, J. Armitage, P. Curtis, R. Abielmona, and E. Petriu, “Navigating complexity: Automating maritime decision-making with temporal transformer-based embeddings and scalable clustering,” 19th IEEE International Systems Conference, 2025, pp. 1-8.

[7] P. Saleh, J. Armitage, P. Curtis, R. Abielmona, and E. Petriu, “Navigating the annotation bottleneck: Active learning for scalable maritime data analytics,” OCEANS 2025 Great Lakes, 2025, in press.

[8] A. Harati-Mokhtari, A. Wall, P. Brooks, and J. Wang, “Automatic Identification System (AIS): Data reliability and human error implications,” Journal of Navigation, vol. 60, no. 3, pp. 373–389, 2007.

[9] M. Riveiro, G. Pallotta, and M. Vespe, “Maritime anomaly detection: A review,” WIREs Data Mining and Knowledge Discovery, vol. 8, no. 5, 2018.

[10] N. M. Gardazi et al., “BERT applications in natural language processing: A review,” Artificial Intelligence Review, vol. 58, 2025.

[11] A. Pal et al., “Chronos: Learning the language of time series,” arXiv preprint arXiv:2403.07815, 2024.

[12] X. Dong, Y. Tian, L. Dai, J. Li, and L. Wan, “A new accurate aircraft trajectory prediction in terminal airspace based on spatio-temporal attention mechanism,” Aerospace, vol. 11, no.9, 2024.

[13] T. T. Um, et al., “Data augmentation of wearable sensor data for parkinson’s disease monitoring using convolutional neural networks,” Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp. 216-220, 2017.

[14] Y. Tang, Y. Zhang, and J. Li, “A time series driven model for early sepsis prediction based on transformer module,” BMC Medical Research Methodology, vol. 24, 2024.

[15] J. Levy-Kramer, k-means-constrained [Computer software]. (2018). https://github.com/joshlk/k-means-constrained.

[16] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” Journal of Computational and Applied Mathematics, vol. 20, pp. 53-65, 1987.

[17] J. Johnson, M. Douze, and H. Jégou, “Billion-scale similarity search with GPUs,” IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535-547, 2021.

[18] J. Wang et al., “Milvus: A purpose-built vector data management system,” Proceedings of the 2021 International Conference on Management of Data, pp. 2614-2627, 2021.

[19] B. Settles, “Active learning literature survey,” Univ. Wisconsin-Madison, 2009.

[20] Z. Wang, X. Fang, X. Tang and C. Wu, “Multi-class active learning by integrating uncertainty and diversity,” in IEEE Access, vol. 6, 2018, pp. 22794-22803.

[21] A. Vaswani et al., “Attention is all you need,” Advances in Neural Information Processing Systems (NeurIPS), 2017.

[22] *Total::Insight*™ Decision Support System. (2025). Larus Technologies.

[23] N. E. Humphries et al., “Environmental context explains Lévy and Brownian movement patterns of marine predators,” Nature, vol. 465, pp. 1066-1069, 2010.

[24] X. Xia, Y. Zhao, and D. Jiang, “Multimodal interaction enhanced representation learning for video emotion recognition,” Frontiers in Neuroscience, vol. 16, 2022.