Efficient Heterogeneous Ensembles for Large-Scale Time Series Classification: A Literature Review and Research Proposal for Combining Hydra and Quant

Urav M. Maniar

Faculty of IT, Monash University

Unit:	FIT5126 – Master's Thesis
Supervisor:	Angus Dempster
Student ID:	30071461

Table of Contents

Introduction

The exponential growth of sensor technologies and automated data collection systems has transformed time series analysis from a specialized technique into a cornerstone of contemporary data science. Embedded within these chronologically ordered measurements lies a wealth of untapped knowledge, crucial for diverse applications ranging from interpreting cardiac rhythms for health diagnoses (Nejedly et al., 2022) to signaling equipment failures through anomalous patterns (Zhao et al., 2019). At the intersection of these domains stands Time Series Classification (TSC), the fundamental challenge of systematically categorizing these temporal sequences into discrete, meaningful classes. This task is inherently complex due to high dimensionality, noise, temporal dependencies, and, for Multivariate Time Series Classification (MTSC), intricate inter-variable relationships (Tong et al., 2022).

Early comprehensive evaluations, notably the "great time series classification bake off" categorized the distinct algorithmic paradigms developed for TSC, highlighting that distance-based, interval-based, shapelet-based, and dictionary-based methods often capture fundamentally different types of information (Bagnall et al., 2017). This development was driven by recognizing the multifaceted nature of time series patterns (e.g., global similarities, local shapes, symbolic motifs), with each paradigm acting as a different analytical "lens" (Bagnall et al., 2017, p. 610). This inherent paradigm diversity underscored the limitation that no single approach consistently dominates across all problem types.

Consequently, heterogeneous ensemble methods emerged as a powerful strategy to leverage these diverse perspectives. By combining classifiers built on different underlying data representations, meta-algorithms like the Collective of Transformation-based Ensembles (COTE) demonstrated superior performance (Bagnall et al., 2017, p. 629). This philosophy culminated in the current state-of-the-art ensemble, HIVE-COTE 2.0 (HC2), which integrates specialists from multiple domains to achieve benchmark accuracy (Middlehurst et al., 2021). However, this peak accuracy comes at a significant computational cost, often involving days or weeks of computation time, limiting practicality in many scenarios (Middlehurst et al., 2021, p. 20, 2024, p. 2012).

Driven partly by this "accuracy-efficiency tension", a recent and significant trend involves developing algorithms that deliver highly competitive accuracy with dramatically reduced computational overhead (Middlehurst et al., 2024, p. 2009). The ROCKET algorithm (Dempster et al., 2020), utilizing random convolutional kernels, marked a key breakthrough in this area, demonstrating that exceptional speed and accuracy were simultaneously achievable. This spurred further development of efficient techniques, pushing the boundaries of practical TSC.

Within this "new wave" of efficient algorithms, two recent advancements stand out, representing leading performance within distinct paradigms. Hydra (Dempster et al., 2023) innovatively bridges dictionary-based methods and convolutional approaches, employing "competing convolutional kernels" organized into groups to efficiently capture multi-scale local pattern frequencies and shapes (Dempster et al., 2023, p. 1786). In contrast, Quant (Dempster et al., 2024) represents a "minimalist interval method," achieving state-of-the-art interval-based accuracy through a deliberate simplification: relying solely on quantile features computed over fixed dyadic intervals to capture segment distributions directly (Dempster et al., 2024, pp. 2377, 2383). Both Hydra and Quant exemplify how high performance can be attained efficiently within their respective algorithmic traditions.

This research focuses specifically on the potential synergy unlocked by ensembling these two modern, efficient algorithms. Hydra and Quant appear theoretically complementary: Hydra excels at identifying which local shapes or patterns are present and dominant via its competing kernels, while Quant summarizes what the statistical distribution of values looks like within various segments using quantiles (Dempster et al., 2023, p. 1786, 2024, p. 2385), offering distinct views of the data.

While sophisticated ensembles like HC2 integrate multiple complex components, and efficient combinations like MultiROCKET+Hydra merge algorithms from similar origins (Middlehurst et al., 2024, p. 2003), the specific potential of a targeted ensemble combining only the efficient Hydra and Quant, leveraging their distinct strengths, remains unexplored. Furthermore, the optimal strategy for integrating their fundamentally different feature representations (whether via simple feature concatenation or a more sophisticated stacking approach (Wolpert, 1992)) is an open question. Such an ensemble holds the promise of achieving accuracy competitive with top-tier methods but with substantially reduced computational overhead. Therefore, the aims of this Literature Review and Proposal are threefold:

1. Critically review the TSC landscape, tracing key paradigms and the drive towards efficiency leading to algorithms like Hydra and Quant.
2. Synthesize literature findings to articulate the rationale and define the research gap regarding ensembling Hydra and Quant.
3. Present a feasible research plan to empirically investigate this gap, evaluating combination strategies and benchmarking performance.

This report proceeds by presenting the substantive literature review, followed by a summary identifying the state-of-the-art and research gap, the proposed research project plan, and concluding remarks.

Substantive Literature Review

This review critically examines the landscape of Time Series Classification (TSC), tracing its evolution from foundational distance-based methods to contemporary state-of-the-art algorithms. Emphasis is placed on understanding the inherent challenges within TSC and how different algorithmic paradigms attempt to address them, often leveraging diverse feature representations (Bagnall et al., 2017, p. 610). This analysis progressively builds the rationale for why ensemble methods have become crucial, setting the stage for evaluating the recent, efficient algorithms Hydra and Quant and exploring the untapped potential of their combination.

The Evolving Landscape of Time Series Classification

Understanding the motivation and potential for modern ensemble techniques requires appreciating the foundational concepts, challenges, and algorithmic diversity within TSC. This section establishes key definitions, discusses the inherent difficulties that drive innovation in the field, and reviews the major families of algorithms that have emerged, highlighting the distinct types of features they capture

Core Concepts and the Rise of Benchmarking

Time Series Classification (TSC) fundamentally involves assigning a discrete class label to an input time series, which is defined as an ordered sequence of observations (Middlehurst et al., 2024, p. 1960). This inherent ordering is the defining characteristic that distinguishes TSC from standard classification tasks, as discriminative features often depend explicitly on the sequence's temporal structure (Middlehurst et al., 2024, p. 1959). Time series data can be broadly categorized based on key characteristics: dimensionality, distinguishing between univariate (UTSC) and multivariate (MTSC) series (Middlehurst et al., 2024, p. 1960; Pasos et al., 2021, p. 403); continuity (discrete vs. continuous observations); and stability (stationary vs. non-stationary statistical properties) (Tong et al., 2022). The central goal of any TSC algorithm is to fit models that effectively capture discriminative information from this ordered data to predict the correct class label (Middlehurst et al., 2024, p. 1959). The establishment of benchmark repositories, notably the UCR Time Series Archive as detailed by Bagnall et al. (2017, p. 607), has been pivotal in enabling rigorous empirical comparison and driving algorithmic advancement in the field.

Intrinsic Challenges and the Motivation for Ensembles

Developing effective TSC algorithms is complicated by several inherent data characteristics, which often necessitate sophisticated modelling approaches, including ensembles. Time series, particularly long UTSC or MTSC instances, represent high-dimensional data where discriminatory patterns can be obscured (Bagnall et al., 2017, p. 646; Pasos et al., 2021, p. 403). MTSC further compounds this, as interactions between dimensions significantly expand the feature space (Pasos et al., 2021, pp. 403, 442). Ensembles combining diverse feature extractors offer a strategy to navigate these vast spaces more effectively than single models. The crucial temporal dependencies within series require methods that respect ordering. While techniques like Dynamic Time Warping (DTW), originally developed for speech recognition (Sakoe & Chiba, 1978), address misalignments (Bagnall et al., 2017, p. 611), capturing the variety of relevant temporal patterns (short-term shapes, long-term trends, seasonality) often benefits from combining algorithms sensitive to different temporal scales or structures, a natural fit for ensembling. For MTSC, modelling inter-variable dependencies is a key challenge, as naive univariate application often fails (Pasos et al., 2021, p. 405). While bespoke MTSC algorithms exist, ensembles provide a flexible framework for combining dimension-dependent and independent information. Furthermore, real-world series frequently contain noise and non-stationarity (Bagnall et al., 2017, p. 617), demanding robust models; ensembles can enhance robustness by averaging out noise or combining models resilient to different data perturbations. Finally, the persistent interpretability vs. accuracy trade-off, where highly accurate models like deep networks or large ensembles (e.g., HIVE-COTE 2.0; Middlehurst et al., 2021) often function as black boxes, a significant barrier in domains like healthcare (Holzinger et al., 2019), motivates exploring ensembles of simpler, potentially more interpretable or efficient base learners.

Foundational Algorithmic Paradigms & Feature Diversity

The challenges inherent in TSC have spurred the development of diverse algorithmic paradigms, each focusing on extracting different types of discriminatory features from time series data. This diversity is fundamental to the success of ensemble methods, which leverage the complementary strengths of different representations. Key paradigms include those based on distance measures, intervals, shapelets, and dictionaries.

Distance-Based Methods

Distance-based methods classify time series by measuring similarity across the entire sequence, often using 'elastic' distance functions to handle temporal distortions (Middlehurst et al., 2024, p. 1970). The k-Nearest Neighbors (k-NN) algorithm, usually 1-NN, is common. Dynamic Time Warping (DTW) (Sakoe & Chiba, 1978) is the cornerstone, using dynamic programming to find optimal alignments despite distortions (Bagnall et al., 2017, pp. 611-612). While powerful, DTW's quadratic complexity $O(m^2)$ per comparison led to research on constraints and alternative measures, though DTW variants remained competitive (Bagnall et al., 2017, p. 612; Wang et al., 2013). Ensemble approaches emerged, such as the Elastic Ensemble (EE) (Lines & Bagnall, 2015) combining multiple distances, and Proximity Forest (PF) (Lucas et al., 2019) integrating distances into tree structures. PF is the current state-of-the-art in this category (Middlehurst et al., 2024, pp. 1972-1973). While effective for global similarity under warping, this paradigm struggles with localized or frequency-based patterns, motivating the need for complementary approaches within broader ensembles.

Interval-Based Methods

Addressing global distance limitations, interval-based algorithms extract features from multiple phase-dependent intervals (contiguous subseries) to capture localized characteristics (Bagnall et al., 2017, p. 617; Middlehurst et al., 2024, p. 1975). Early methods like Time Series Forest (TSF) (Deng et al., 2013) used simple statistics from random intervals within tree ensembles. Subsequent developments incorporated richer statistical and spectral features, supervised interval selection, and focused distributional summaries like quantiles (e.g., QUANT - Dempster et al., 2024). These methods capture localized properties distinct from global distance or shape features, adding valuable diversity for ensembles. The current efficient state-of-the-art in this category, QUANT, achieves high accuracy using only quantile features over fixed intervals (Dempster et al., 2024; Middlehurst et al., 2024, p. 1979).

Shapelet-Based Methods

Shapelet-based algorithms identify short, phase-independent subsequences ('shapelets') that are maximally discriminative, classifying based on minimum distance to these shapes (Bagnall et al., 2017, p. 622; Middlehurst et al., 2024, p. 1980). While early methods embedding the expensive $O(n^2m^4)$ search in trees were slow (Ye & Keogh, 2011), the Shapelet Transform

Classifier (STC) (Hills et al., 2014) decoupled discovery (using randomized search for scalability) from classification, transforming data based on shapelet distances before applying a standard classifier (Bostrom & Bagnall, 2015). Recent advances like the Random Dilated Shapelet Transform (RDST) (Guillaume et al., 2022) incorporate dilation and randomization with efficient linear classifiers, achieving state-of-the-art shapelet-based performance (Middlehurst et al., 2024, p. 1983). Shapelets capture characteristic local shapes irrespective of position, offering features orthogonal to global distance or interval statistics, thus contributing unique perspectives to heterogeneous ensembles.

Dictionary-Based Methods

Dictionary-based classifiers transform time series into discrete symbolic sequences ('words') and classify based on word frequency distributions, akin to 'bag-of-words' (Bagnall et al., 2017, pp. 626-627; Middlehurst et al., 2024, p. 1984). This approach excels when pattern frequency, not just presence, is key (Bagnall et al., 2017, p. 626). Typically, sliding windows are applied, subsequences approximated (e.g., via PAA or DFT), and then discretized into symbols using methods like SAX (Lin et al., 2007) or SFA (Schäfer & Högqvist, 2012). The influential BOSS algorithm (Schäfer, 2015) utilized SFA and ensembling. Subsequent work focused on improving efficiency, incorporating spatial information, and adding supervised discretization/feature selection. The current state-of-the-art, WEASEL 2.0 (Schäfer & Leser, 2023), integrates dilation and randomization with a linear classifier for high accuracy (Middlehurst et al., 2024, pp. 1988-1990). By capturing pattern recurrence, dictionary methods offer complementary features for ensemble models.

The New Wave: Efficiency and Accuracy

While foundational paradigms established diverse feature extraction techniques, many state-of-the-art algorithms derived from them suffered from high computational complexity (e.g., original STC, BOSS, EE). A significant recent trend, spurred partly by the success of deep learning principles and the need for scalability, involves developing algorithms that achieve competitive or superior accuracy with drastically improved efficiency. This "new wave" includes convolutional kernel methods and highly optimized interval and dictionary approaches.

Convolutional Approaches: The ROCKET Family

Convolutional or kernel-based methods apply filters (kernels) across a time series using a sliding dot-product operation, creating activation maps from which features are extracted, often via pooling (Middlehurst et al., 2024, p. 1991). While inspired by Convolutional Neural Networks (CNNs), this paradigm often uses fixed or randomly generated kernels coupled with a separate, typically linear, classifier. The Random Convolutional Kernel Transform (ROCKET) marked a significant breakthrough, demonstrating that a large number (e.g., 10,000) of random convolutional kernels—with randomized length, weights, bias, dilation, and padding—could achieve state-of-the-art accuracy with exceptional speed (Dempster et al., 2020, p. 1455). ROCKET extracts two features from each kernel's activation map: the global maximum value and the Proportion of Positive Values (PPV), a novel pooling operator capturing the prevalence of a pattern match relative to a bias threshold (Dempster et al., 2020, p. 1463). This high-dimensional feature set (20,000 features) proved highly effective when paired with a simple Ridge Regression classifier (Dempster et al., 2020, p. 1464). Building on ROCKET's success, MiniROCKET achieved comparable accuracy but with significantly greater speed (up to 75x faster) by using a small, fixed set of 84 deterministic kernels (derived from specific weight patterns) and optimizing the PPV calculation (Dempster et al., 2021, p. 248). MultiROCKET further refined this by applying the MiniROCKET kernels to both the original series and its first-order difference, and introducing three additional pooling operators alongside PPV, resulting in accuracy competitive with the most complex ensembles (Tan et al., 2022, p. 1625). The ROCKET family demonstrates the remarkable effectiveness of diverse convolutional features, even when randomly generated or derived from a small deterministic set, combined with efficient linear classifiers.

Hydra: Bridging Convolutions and Dictionaries

Hydra (HYbrid Dictionary-ROCKET Architecture) represents a deliberate attempt to merge the strengths of the ROCKET family's random convolutional kernels with the pattern-counting philosophy of dictionary-based methods (Dempster et al., 2023, p. 1779). It aims to provide an efficient yet powerful dictionary-like representation using these kernels.

Core Mechanism & Feature Extraction

Hydra leverages random convolutional kernels, similar to ROCKET, but introduces a crucial organizational structure: the kernels are arranged into g distinct groups, each containing k kernels (Dempster et al., 2023, p. 1781). For each dilation value applied, the input time series (and typically its first-order difference) is convolved with all kernels. Within each group and at each time point, a competition determines the kernel with the maximum response (largest dot product magnitude) and, optionally, the minimum response (Dempster et al., 2023, p. 1788). Features are then derived by counting these "winning" kernels across the series. Two main counting methods are used: 'hard counting' increments a counter for the winning kernel's index (argmax), effectively tallying how often each kernel best represents the input locally; 'soft counting' accumulates the actual maximum (or minimum) response value for the winning kernel (max/min), capturing the strength of the best match (Dempster et al., 2023, p. 1789). The final feature vector concatenates these counts across all groups, dilations, counting methods (max/min, hard/soft), and series representations (raw/differenced).

Strengths, Weaknesses, and Captured Patterns

Hydra's hybrid design offers several theoretical strengths. It inherits the efficiency of random convolutions from ROCKET (Dempster et al., 2023, p. 1780) while its grouping mechanism provides a structured way to interpret kernel activations, moving beyond simple pooling towards quantifying the relative dominance of detected local shapes or motifs (Dempster et al., 2023, p. 1781). By processing both raw and differenced series and using multiple dilations, it captures multi-scale temporal dynamics and trend information (Dempster et al., 2023, p. 1787). However, the randomness of the kernels makes direct feature interpretation challenging (Dempster et al., 2023, p. 1792). While the competitive grouping aims to structure the feature space, potential redundancy among the many random kernels might still exist. Its primary focus on local patterns means it might be less adept than interval-based methods at capturing purely distributional characteristics or global statistics not reflected in local shapes. Similarly, frequency information is captured only indirectly through kernel shapes, potentially less effectively than dedicated spectral methods like SFA used in BOSS (Schäfer, 2015). Overall, Hydra excels at efficiently identifying the prevalence and strength of diverse, multi-scale local patterns.

Quant: Minimalist Interval-Based Power

Representing the interval-based paradigm within the new wave of efficient classifiers, Quant achieves state-of-the-art accuracy for its category through a deliberately minimalist design (Dempster et al., 2024, p. 2377). It eschews complex feature engineering and interval selection, focusing instead on robust distributional summaries over fixed intervals.

Core Mechanism & Feature Extraction

Quant processes multiple representations of the input series: the raw series, smoothed first difference, second difference, and the magnitude spectrum from a Discrete Fourier Transform (Dempster et al., 2024, p. 2383). For each representation, it extracts features from a fixed set of dyadic intervals (progressively halving the series length) and corresponding shifted intervals (Dempster et al., 2024, p. 2384, Fig. 3). The core innovation is the use of quantiles as the sole feature type extracted from each interval. By default, it computes m/4 evenly spaced quantiles for an interval of length m (Dempster et al., 2024, p. 2385). To capture the distribution relative to the local mean and add robustness to level shifts, the interval mean is subtracted before computing every second quantile (Dempster et al., 2024, p. 2386). All quantile features from all intervals across all four representations are concatenated. This feature vector is then classified using a standard Extremely Randomized Trees (ExtraTrees) classifier (Dempster et al., 2024, p. 2386).

Strengths, Weaknesses, and Captured Patterns

Quant's primary strength lies in its exceptional speed and scalability, stemming from the efficiency of quantile computation and the use of fixed intervals (Dempster et al., 2024, p. 2378). Quantiles provide a general and robust summary of interval distributions, capturing central tendency, spread, and tail behavior while being less sensitive to outliers than mean or variance (Dempster et al., 2024, p. 2387). Explicitly processing multiple representations (raw, differences, frequency) allows it to capture diverse characteristics related to level, change, and periodicity within intervals (Dempster et al., 2024, p. 2383). However, its reliance on statistical summaries means it inherently loses fine-grained intra-interval sequence information and local shape details (Dempster et al., 2024, p. 2385). The fixed dyadic intervals might not optimally align with meaningful events in all series (Dempster et al., 2024, p. 2386). Furthermore, it generates a potentially large feature set and relies entirely on the downstream classifier (ExtraTrees) for implicit feature selection (Dempster et al., 2024, p. 2387). Quant excels at efficiently classifying based on the statistical profile of series segments across different data views.

Ensembling Strategies in TSC

Given that different TSC algorithms capture distinct feature types, combining multiple classifiers into an ensemble is a well-established strategy for improving accuracy and robustness. The most successful ensembles in TSC, particularly heterogeneous ones, demonstrate the power of integrating diverse perspectives on the data.

Heterogeneous Ensembles: The HIVE-COTE Benchmark

The principle that combining diverse representations leads to superior performance is best exemplified by the Hierarchical Vote Collective of Transformation-based Ensembles (HIVE-COTE) family (Lines et al., 2018). HIVE-COTE acts as a meta-ensemble, integrating specialist classifiers built on different data transformations. The current state-of-the-art version, HIVE-COTE 2.0 (HC2), combines classifiers from shapelet (STC), dictionary (TDE), interval (DrCIF), and convolutional (Arsenal - an ensemble of ROCKET models) domains (Middlehurst et al., 2021, p. 6). HC2 employs the Cross-validation Accuracy Weighted Probabilistic Ensemble (CAWPE) mechanism, which weights the probabilistic outputs of each component based on an estimate of its accuracy derived from the training data (Large et al., 2019; Middlehurst et al., 2021, p. 5). While HC2 achieves benchmark accuracy on UCR/UEA archives (Middlehurst et al., 2021, p. 2, 2024, p. 2005), its complexity and computational cost, stemming from integrating multiple complex components, remain significant (Middlehurst et al., 2021, p. 20). This motivates the exploration of simpler heterogeneous ensembles using more efficient base learners.

Combining Efficient Algorithms: The Hydra & Quant Case

The demonstrated complementarity of Hydra (capturing local, multi-scale shapes and patterns) and Quant (capturing interval-based distributional statistics across multiple data views) makes them prime candidates for a potentially powerful yet efficient heterogeneous ensemble, mirroring the HIVE-COTE philosophy with modern components.

Analyzing Feature Complementarity (Hydra vs. Quant)

As established in Sections 2.2.2.2 and 2.2.3.2, Hydra and Quant probe fundamentally different aspects of time series data. Hydra focuses on identifying the presence and relative dominance of specific local shapes and patterns via its competing convolutional kernels (Dempster et al., 2023, p. 1781), making it sensitive to morphology. Quant, conversely, summarizes the statistical distribution of values within fixed intervals using quantiles, capturing characteristics like central tendency, spread, and skewness independent of the exact intra-interval sequence (Dempster et al., 2024, p. 2387). This orthogonality suggests their features provide complementary views: Hydra captures 'what shapes are present and where/how often', while Quant captures 'what the value distributions look like in different segments'. Combining these views should theoretically yield a more comprehensive data representation than either algorithm alone, potentially leading to improved classification performance, particularly on complex datasets where both local patterns and distributional shifts are relevant.

Evaluating Combination Methods: Concatenation vs. Stacking

Two primary strategies exist for combining Hydra and Quant. Feature concatenation involves simply merging the feature vectors produced by both algorithms into a single, larger vector, which is then used to train a classifier (e.g., Ridge Regression). This approach preserves all original feature information but creates a highly heterogeneous and potentially very high-dimensional space, which might challenge simpler linear classifiers. The successful use of concatenation in MultiRocket+Hydra provides some precedent, though both components stem from the convolutional paradigm (Dempster et al., 2023, p. 1796). Stacking (stacked generalization) offers an alternative where Hydra and Quant are trained independently, and their predictions (e.g., class probabilities derived via cross-validation) become features for a second-level meta-classifier (Wolpert, 1992). This approach inherently handles feature heterogeneity by operating in the probability space and explicitly learns how to combine the 'judgments' of the base models. It conceptually mirrors HIVE-COTE's CAWPE mechanism (Middlehurst et al., 2021, p. 5), but risks information loss as the meta-learner doesn't see the original features. Given the distinct nature of Hydra's pattern-based features and Quant's distributional features, stacking appears theoretically well-suited to optimally weigh their diverse contributions, though empirical validation comparing both methods is necessary.

Summary of the State of the Art

The field of Time Series Classification (TSC) has matured significantly, moving beyond foundational methods like 1-NN DTW towards more sophisticated and accurate techniques. The literature reveals a rich tapestry of algorithmic paradigms, each designed to capture distinct types of discriminatory information embedded within time series data. Distance-based methods excel at handling global similarities under temporal warping (Bagnall et al., 2017, pp. 611–612), while interval-based approaches focus on localized statistical or spectral properties within phase-dependent segments (Middlehurst et al., 2024, p. 1975). Shapelet-based algorithms identify characteristic, phase-independent local patterns (Bagnall et al., 2017, p. 622), and dictionary-based methods quantify the frequency of recurring symbolic motifs (Bagnall et al., 2017, pp. 626–627). More recently, convolutional approaches, exemplified by the ROCKET family (Dempster et al., 2020, 2021; Tan et al., 2022), leverage random or deterministic kernels to efficiently extract diverse features, and deep learning models like H-InceptionTime automatically learn hierarchical representations (Middlehurst et al., 2024, p. 2001).

A key finding, consistently reinforced by comprehensive benchmarks (Bagnall et al., 2017; Middlehurst et al., 2024), is that no single paradigm dominates across all types of TSC problems. This inherent feature diversity has driven the success of heterogeneous ensemble methods. The HIVE-COTE family, culminating in HIVE-COTE 2.0 (Middlehurst et al., 2021), represents the current state-of-the-art in accuracy by explicitly combining specialist classifiers from multiple domains (shapelet, dictionary, interval, convolution) using a sophisticated weighting scheme (Middlehurst et al., 2021, p. 6). However, this benchmark accuracy comes at a significant computational cost, limiting its practicality in resource-constrained or time-sensitive applications (Middlehurst et al., 2021, p. 20).

Concurrently, a "new wave" of algorithms has emerged, prioritizing computational efficiency while maintaining high accuracy. Within this wave, Hydra (Dempster et al., 2023) and Quant (Dempster et al., 2024) stand out as highly competitive representatives of different paradigms. Hydra effectively bridges convolutional and dictionary approaches, efficiently capturing multi-scale local patterns using competing random kernels (Dempster et al., 2023, p. 1781). Quant offers a minimalist yet powerful interval-based approach, demonstrating state-of-the-art performance within its category by focusing solely on quantile features across multiple data representations (Dempster et al., 2024, p. 2377). Crucially, Hydra and Quant appear theoretically complementary, with Hydra focusing on local morphology and Quant on interval statistics (Sections 2.2.2.2 & 2.2.3.2).

This confluence defines the current research gap: while the power of heterogeneous ensembling is proven by HIVE-COTE 2.0, and efficient, complementary algorithms like Hydra and Quant exist, the specific potential of combining just these two efficient algorithms to achieve high accuracy has not been explored. Such an ensemble could potentially offer a significant portion of HIVE-COTE 2.0's accuracy but with drastically reduced computational overhead. Furthermore, the optimal method for integrating their distinct feature types—whether through simple feature concatenation or a more sophisticated stacking approach—remains an open question (Section 2.3.2.2). Therefore, a systematic investigation into ensembling Hydra and Quant, evaluating different combination strategies, is needed to determine if a more efficient yet highly accurate heterogeneous ensemble can be constructed, pushing the state-of-the-art boundary for practical TSC.

Research Project Plan

Building upon the literature review and the identified research gap concerning the unexplored potential of combining the efficient Hydra and Quant algorithms (Section 3), this section details the plan for empirically investigating their synergistic combination. Emphasizing the need to evaluate performance on more challenging, large-scale problems, the central aim is to determine if a Hydra+Quant ensemble can achieve high classification accuracy and efficiency on contemporary, large datasets, thereby addressing the limitations of benchmarks focused solely on smaller datasets and the accuracy-efficiency tension prevalent in Time Series Classification (TSC).

Research Aim and Questions

The primary aim of this research project is: *To investigate the effectiveness and efficiency of ensembling the Hydra and Quant algorithms for Time Series Classification on large-scale datasets, comparing different combination strategies and benchmarking against relevant state-of-the-art (SOTA) methods.*

This aim will be addressed through the following specific research questions (RQs):

RQ1: Which combination strategy (or strategies) yields the optimal trade-off between classification accuracy and computational efficiency for a Hydra+Quant ensemble, considering methods beyond basic concatenation and stacking?

RQ2: How does the performance (accuracy and efficiency) of the optimized Hydra+Quant ensemble compare against individual Hydra, Quant, and selected efficient SOTA classifiers (e.g., MultiROCKET) on representative large-scale TSC benchmarks?

RQ3: How does the performance of the Hydra+Quant ensemble compare to the high-accuracy SOTA represented by HIVE-COTE 2.0 (using reported results), specifically regarding the accuracy-efficiency trade-off?

Proposed Methodology

To systematically address the research questions, the following methodology will be employed, adhering to established practices within the TSC research community and ensuring feasibility within the project timeframe.

Datasets and Benchmarking Framework

The primary experimental testbed will be the MONSTER (Monash Scalable Time Series Evaluation Repository) archive (Dempster et al., 2025). This choice is driven by the need to evaluate algorithms on datasets that better reflect the scale and complexity of many real-world TSC problems and to address the limitations of traditional benchmarks like UCR, which primarily consist of smaller datasets where variance minimization often dominates performance. The MONSTER archive provides a crucial platform for assessing the scalability and large-data performance central to this project's aims.

Given the large size of the MONSTER archive (29 datasets ranging up to ~59M instances) and the 6-month project timeframe, a representative subset of MONSTER datasets (aiming for approximately 5-10 datasets) will be selected for the core experiments. This subset will aim to cover diverse data characteristics, including:

- Different data types (e.g., Audio, Satellite Image Time Series, EEG, HAR).
- Varying dataset sizes (number of instances, e.g., ranging from tens of thousands to millions).
- Varying time series lengths and numbers of channels (univariate and multivariate).

The evaluation will utilize the pre-defined 5-fold cross-validation splits provided with the MONSTER datasets. Using these splits is critical as it ensures consistency, allows for direct comparison with baseline results reported in the MONSTER paper and future work using this archive, and often accounts for inherent data structures (like subject or location dependencies) that random splitting might ignore. Performance metrics will be averaged across these 5 folds for robust evaluation.

Implementation and Ensemble Strategies

The core algorithms under investigation are Hydra and Quant, chosen for their efficiency and theoretically complementary approaches (local pattern detection vs. interval statistics). Implementations will primarily be sourced from established Python toolkits like `aeon`, ensuring reliance on validated, reproducible codebases. Original author implementations may be adapted if necessary.

A central part of this research (RQ1) involves exploring effective ensemble strategies for combining Hydra and Quant. Simple feature concatenation, where feature vectors from both algorithms are merged before training a linear classifier (e.g., Ridge Regression), will serve as a baseline. Stacking provides another baseline, where Hydra and Quant are trained independently, and their out-of-fold predictions (class probabilities) become meta-features for a second-level classifier (e.g., Logistic Regression).

However, acknowledging the lack of prior work combining these specific algorithms and the potential limitations of generic strategies, further hybrid feature/prediction integration methods will be investigated, following supervisor suggestions. This includes exploring a Hydra-Informed Quant approach, where Quant's classifier (ExtraTrees) input is augmented with Hydra's probability predictions, and conversely, a Quant-Informed Hydra where Hydra's classifier (Ridge Regression) input is augmented with Quant's predictions. Furthermore, given that Hydra's features can be individually weak, Hydra feature aggregation strategies will be explored. This might involve training ensembles of smaller "mini-Hydras" and combining their aggregated features or predictions with Quant's output, potentially via concatenation or stacking variants. This multi-faceted investigation is motivated by the need to empirically determine how best to leverage the distinct information captured by Hydra and Quant in the absence of established methods for this specific pairing, recognizing that the 6-month scope necessitates focused exploration rather than exhaustive testing of all possibilities.

To evaluate the performance of the developed ensembles (RQ2, RQ3), results will be benchmarked against several key algorithms. The individual Hydra and Quant algorithms, run with their standard configurations (e.g., Ridge Classifier for Hydra, ExtraTrees for Quant), will provide essential baselines to quantify the ensemble benefit. MultiROCKET will be included as a representative efficient SOTA classifier from the convolutional paradigm, offering a comparison point for accuracy and speed. Finally, HIVE-COTE 2.0 will serve as the high-accuracy SOTA benchmark; due to its significant computational demands, its performance will primarily be referenced from reliable reported results on the selected MONSTER datasets rather than being extensively re-run within this project.

Evaluation Metrics and Statistical Analysis

Performance will be comprehensively evaluated using standard metrics to address the accuracy-efficiency trade-off central to the research questions. The primary metrics include Classification Accuracy as the main indicator of predictive correctness, Negative Log Loss (NLL) to assess the quality of probabilistic predictions (especially relevant for stacking/hybrid methods), and Computational Time (wall-clock time for training and prediction, measured separately) to quantify efficiency.

As a secondary metric, the F1-Score may be reported alongside accuracy, particularly for datasets within the selected MONSTER subset exhibiting significant class imbalance, providing a more nuanced view of performance across classes. For comparison and analysis:

- Results will be averaged across the 5 pre-defined MONSTER CV folds for robustness.
- Performance will be compared using tables summarizing average metrics (Accuracy, NLL, Train Time, Predict Time) across the selected datasets for each algorithm and ensemble configuration.
- Pairwise comparisons (e.g., scatter plots, win/draw/loss counts) will be used to visualize relative performance between ensemble strategies and baselines.
- The analysis will focus on identifying the ensemble strategy offering the best accuracy-efficiency profile (RQ1) and quantifying the performance relative to individual components and SOTA benchmarks (RQ2, RQ3) on large-scale data. While formal statistical tests (e.g., Wilcoxon signed rank with Holm correction) might be used for key comparisons, the emphasis will be on practical significance and trade-offs.

Justification and Expected Contributions

This research is justified by its direct response to the identified gap in the literature: the lack of empirical investigation into ensembling the efficient, theoretically complementary Hydra and Quant algorithms, particularly on large-scale data. The focus on the MONSTER archive addresses the need to evaluate algorithms under conditions relevant to modern data scales and the accuracy-efficiency tension, moving beyond traditional benchmarks. The proposal to explore multiple ensemble strategies, including hybrid approaches beyond basic concatenation/stacking, is motivated by supervisor feedback and the lack of prior work, making it a novel investigation. The selected baselines provide essential comparisons against individual components, efficient SOTA, and high-accuracy SOTA. The chosen evaluation metrics directly measure effectiveness and efficiency. The phased timeline and reliance on existing toolkits ensure feasibility within the 6-month Master's project scope. Overall, this plan is designed to be persuasive by clearly linking the identified gap to a specific, appropriate, and feasible methodology aimed at generating novel insights. The expected contributions of this project to the field of Time Series Classification include:

1. Empirical evidence on the synergistic potential (or lack thereof) of combining Hydra and Quant.
2. A comparative analysis of feature concatenation, stacking, and novel hybrid integration strategies for these specific algorithms on large-scale data.
3. Performance benchmarks of potentially novel, efficient heterogeneous ensembles (Hydra+Quant variants) against state-of-the-art methods on challenging, large-scale datasets from the MONSTER archive.
4. Insights into the accuracy-efficiency trade-offs when constructing ensembles with modern, efficient TSC components, providing guidance for practitioners facing resource constraints.
5. An initial exploration into the uncharted space of combining Hydra and Quant, potentially highlighting promising avenues for future research.

Project Timeline

A high-level phased timeline provides a sensible structure for the 6-month research period, detailed in Table 1.

Ethics Considerations

This research will primarily utilize publicly available benchmark datasets, specifically from the MONSTER archive and potentially referencing UCR archive data. These archives contain anonymized or non-personal data standardly used for algorithmic research in the TSC community. Therefore, the project does not involve direct human interaction or the collection of sensitive personal data. Formal ethics committee approval is not expected to be required. All research activities will be conducted in strict compliance with Monash University's research integrity standards and data usage policies.

Conclusion

This Literature Review and Proposal has explored the dynamic field of Time Series Classification (TSC), highlighting the critical tension between achieving high classification accuracy and maintaining computational efficiency, especially in the face of increasingly large datasets. The review traced the evolution from foundational methods to diverse algorithmic paradigms (distance, interval, shapelet, dictionary, convolution-based), noting that no single approach dominates all problem types. This diversity motivated the development of heterogeneous ensembles like HIVE-COTE 2.0, which achieves state-of-the-art accuracy but at significant computational cost.

A "new wave" of efficient algorithms, including the convolution/dictionary-based Hydra and the minimalist interval-based Quant, offers highly competitive performance with reduced overhead. Their theoretical complementarity—Hydra focusing on local patterns and Quant on interval distributions—presents an unexplored opportunity for creating a powerful yet efficient ensemble. This forms the central research gap addressed by this proposal.

The proposed research project will empirically investigate the synergy of ensembling Hydra and Quant, focusing on large-scale datasets from the MONSTER archive. It will compare various combination strategies, moving beyond simple concatenation and stacking to explore novel hybrid approaches. Performance will be rigorously benchmarked against individual components and relevant state-of-the-art methods, evaluating both accuracy and computational efficiency. This research aims to contribute valuable insights into constructing effective and practical ensembles for modern TSC challenges, potentially offering a highly accurate solution with significantly lower computational demands than current complex meta-ensembles.

Tables

Table 1

Proposed Project Timeline

Phase	Duration	Key Tasks	Key Milestones
Phase 1	1 Month	Finalize MONSTER dataset subset selection. Set up computational environment (aeon, libraries); Implement/verify Hydra, Quant, MultiROCKET. Run baseline experiments. Collect initial accuracy/timing data. Refine parameters if needed.	Dataset subset finalized. Core algorithms verified. Baseline performance established.
Phase 2	2 Months	Implement concatenation, stacking, and 1-2 selected hybrid ensemble strategies; Run main ensemble experiments across all datasets/folds; Collect detailed accuracy and timing data.	Core ensemble strategies tested. Main experimental results collected.
Phase 3	3 Months	Analyze collected data (RQ1, RQ2, RQ3); Incorporate reported HC2 results; Generate plots/tables; Interpret findings; If time permits, explore additional hybrid strategies/sensitivities; Write Master's thesis.	Data analysis complete. Thesis draft written and revised. Final submission.

This timeline provides structure while allowing some flexibility for analysis and potential minor exploratory experiments in the later stages.

References

Bagnall, A., Lines, J., Bostrom, A., Large, J., & Keogh, E. (2017). The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery, 31(3)*, 606–660. https://doi.org/10.1007/s10618-016-0483-9

Bostrom, A., & Bagnall, A. (2015). Binary Shapelet Transform for Multiclass Time Series Classification. In S. Madria & T. Hara (Eds.), *Big Data Analytics and Knowledge Discovery* (pp. 257–269). Springer International Publishing. https://doi.org/10.1007/978-3-319-22729-0_20

Cabello, N., Naghizade, E., Qi, J., & Kulik, L. (2020). Fast and Accurate Time Series Classification Through Supervised Interval Search. 2020 *IEEE International Conference on Data Mining (ICDM)*, 948–953. https://doi.org/10.1109/ICDM50108.2020.00107

Chen, L., & Ng, R. (2004). - On The Marriage of Lp-norms and Edit Distance. In M. A. Nascimento, M. T. Özsu, D. Kossmann, R. J. Miller, J. A. Blakeley, & B. Schiefer (Eds.), *Proceedings 2004 VLDB Conference* (pp. 792–803). Morgan Kaufmann. https://doi.org/10.1016/B978-012088469-8.50070-X

Dempster, A., Foumani, N. M., Tan, C. W., Miller, L., Mishra, A., Salehi, M., Pelletier, C., Schmidt, D. F., & Webb, G. I. (2025). *MONSTER: Monash Scalable Time Series Evaluation Repository* (No. arXiv:2502.15122). arXiv. https://doi.org/10.48550/arXiv.2502.15122

Dempster, A., Petitjean, F., & Webb, G. I. (2020). ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34(5), 1454–1495. https://doi.org/10.1007/s10618-020-00701-z

Dempster, A., Schmidt, D. F., & Webb, G. I. (2021). MINIROCKET: A Very Fast (Almost) Deterministic Transform for Time Series Classification. *arXiv.Org*. https://doi.org/10.48550/arxiv.2012.08791

Dempster, A., Schmidt, D. F., & Webb, G. I. (2023). Hydra: Competing convolutional kernels for fast and accurate time series classification. *Data Mining and Knowledge Discovery*, 37(5), 1779–1805. https://doi.org/10.1007/s10618-023-00939-3

Dempster, A., Schmidt, D. F., & Webb, G. I. (2024). quant: A minimalist interval method for time series classification. *Data Mining and Knowledge Discovery*, 38(4), 2377–2402. https://doi.org/10.1007/s10618-024-01036-9

Deng, H., Runger, G., Tuv, E., & Vladimir, M. (2013). A time series forest for classification and feature extraction. *Information Sciences*, 239, 142–153. https://doi.org/10.1016/j.ins.2013.02.030

Flynn, M., Large, J., & Bagnall, T. (2019). The Contract Random Interval Spectral Ensemble (c-RISE): The Effect of Contracting a Classifier on Accuracy. In H. Pérez García, L. Sánchez González, M. Castejón Limas, H. Quintián Pardo, & E. Corchado Rodríguez (Eds.), *Hybrid Artificial Intelligent Systems* (pp. 381–392). Springer International Publishing. https://doi.org/10.1007/978-3-030-29859-3_33

Guillaume, A., Vrain, C., & Elloumi, W. (2022). Random Dilated Shapelet Transform: A New Approach for Time Series Shapelets. In M. El Yacoubi, E. Granger, P. C. Yuen, U. Pal, & N. Vincent (Eds.), *Pattern Recognition and Artificial Intelligence* (pp. 653–664). Springer International Publishing. https://doi.org/10.1007/978-3-031-09037-0_53

Hills, J., Lines, J., Baranauskas, E., Mapp, J., & Bagnall, A. (2014). Classification of time series by shapelet transformation. *Data Mining and Knowledge Discovery*, 28(4), 851–881. https://doi.org/10.1007/s10618-013-0322-1

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, 9(4), e1312. https://doi.org/10.1002/widm.1312

Large, J., Bagnall, A., Malinowski, S., & Tavenard, R. (2019). On time series classification with dictionary-based classifiers. *Intelligent Data Analysis*, 23(5), 1073–1089. https://doi.org/10.3233/IDA-184333

Lin, J., Keogh, E., Wei, L., & Lonardi, S. (2007). Experiencing SAX: A novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, 15(2), 107–144. https://doi.org/10.1007/s10618-007-0064-z

Lin, J., Khade, R., & Li, Y. (2012). Rotation-invariant similarity in time series using bag-of-patterns representation. *Journal of Intelligent Information Systems*, 39(2), 287–315. https://doi.org/10.1007/s10844-012-0196-5

Lines, J., & Bagnall, A. (2015). Time series classification with ensembles of elastic distance measures. *Data Mining and Knowledge Discovery*, 29(3), 565–592. https://doi.org/10.1007/s10618-014-0361-2

Lines, J., Taylor, S., & Bagnall, A. (2018). Time Series Classification with HIVE-COTE: The Hierarchical Vote Collective of Transformation-Based Ensembles. *ACM Transactions on Knowledge Discovery from Data*, 12(5), 1–35. https://doi.org/10.1145/3182382

Lucas, B., Shifaz, A., Pelletier, C., O'Neill, L., Zaidi, N., Goethals, B., Petitjean, F., & Webb, G. I. (2019). Proximity Forest: An effective and scalable distance-based classifier for time series. *Data Mining and Knowledge Discovery*, 33(3), 607–635. https://doi.org/10.1007/s10618-019-00617-3

Middlehurst, M., Large, J., & Bagnall, A. (2020). The Canonical Interval Forest (CIF) Classifier for Time Series Classification. *arXiv.Org*. https://doi.org/10.1109/BigData50022.2020.9378424

Middlehurst, M., Large, J., Flynn, M., Lines, J., Bostrom, A., & Bagnall, A. (2021). HIVE-COTE 2.0: A new meta ensemble for time series classification. *Machine Learning, 110*(11–12), 3211–3243. https://doi.org/10.1007/s10994-021-06057-9

Middlehurst, M., Schäfer, P., & Bagnall, A. (2024). Bake off redux: A review and experimental evaluation of recent time series classification algorithms. *Data Mining and Knowledge Discovery*, 38(4), 1958–2031. https://doi.org/10.1007/s10618-024-01022-1

Middlehurst, M., Vickers, W., & Bagnall, A. (2019). Scalable Dictionary Classifiers for Time Series Classification. In H. Yin, D. Camacho, P. Tino, A. J. Tallón-Ballesteros, R. Menezes, & R. Allmendinger (Eds.), *Intelligent Data Engineering and Automated Learning – IDEAL 2019* (pp. 11–19). Springer International Publishing. https://doi.org/10.1007/978-3-030-33607-3_2

Nejedly, P., Ivora, A., Viscor, I., Koscova, Z., Smisek, R., Jurak, P., & Plesinger, F. (2022). Classification of ECG using ensemble of residual CNNs with or without attention mechanism. *Physiological Measurement, 43*(4), 044001. https://doi.org/10.1088/1361-6579/ac647c

Pasos, R. A., Flynn, M., Large, J., Middlehurst, M., & Bagnall, A. (2021). The great multivariate time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 35(2), 401–449. https://doi.org/10.1007/s10618-020-00727-3

Ratanamahatana, C. A., & Keogh, E. (2005). Three Myths about Dynamic Time Warping Data Mining. *Society for Industrial and Applied Mathematics. Proceedings of the SIAM International Conference on Data Mining*, 506–510. https://www.proquest.com/docview/942758626/abstract/BB151AFB7B60483BPQ/1

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1), 43–49. https://doi.org/10.1109/TASSP.1978.1163055

Schäfer, P. (2015). The BOSS is concerned with time series classification in the presence of noise. *Data Mining and Knowledge Discovery*, 29(6), 1505–1530. https://doi.org/10.1007/s10618-014-0377-7

Schäfer, P., & Högqvist, M. (2012). SFA: A symbolic fourier approximation and index for similarity search in high dimensional datasets. *Proceedings of the 15th International Conference on Extending Database Technology*, 516–527. https://doi.org/10.1145/2247596.2247656

Schäfer, P., & Leser, U. (2017). Fast and Accurate Time Series Classification with WEASEL. *arXiv.Org*. https://doi.org/10.1145/3132847.3132980

Schäfer, P., & Leser, U. (2023). WEASEL 2.0: A random dilated dictionary transform for fast, accurate and memory constrained time series classification. *Machine Learning*, 112(12), 4763–4788. https://doi.org/10.1007/s10994-023-06395-w

Senin, P., & Malinchik, S. (2013). SAX-VSM: Interpretable Time Series Classification Using SAX and Vector Space Model. *2013 IEEE 13th International Conference on Data Mining*, 1175–1180. https://doi.org/10.1109/ICDM.2013.52

Tan, C. W., Dempster, A., Bergmeir, C., & Webb, G. I. (2022). MultiRocket: Multiple pooling operators and transformations for fast and effective time series classification. *Data Mining and Knowledge Discovery, 36*(5), 1623–1646. https://doi.org/10.1007/s10618-022-00844-1

Tong, Y., Liu, J., Yu, L., Zhang, L., Sun, L., Li, W., Ning, X., Xu, J., Qin, H., & Cai, Q. (2022). Technology investigation on time series classification and prediction. *PeerJ Computer Science*. https://doi.org/10.7717/peerj-cs.982

Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., & Keogh, E. (2013). Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, 26(2), 275–309. https://doi.org/10.1007/s10618-012-0250-5

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Ye, L., & Keogh, E. (2011). Time series shapelets: A novel technique that allows accurate, interpretable and fast classification. *Data Mining and Knowledge Discovery*, 22(1–2), 149–182. https://doi.org/10.1007/s10618-010-0179-5

Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2019). Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing*, 115, 213–237. https://doi.org/10.1016/j.ymssp.2018.05.050