DYNAMIC TIME WARPING AND DUAL-ATTENTION NETWORK FOR MULTIVARIATE TIME SERIES CLASSIFICATION PROBLEMS

LIM YEE LYN

UNIVERSITI KEBANGSAAN MALAYSIA

DYNAMIC TIME WARPING AND DUAL-ATTENTION NETWORK FOR MULTIVARIATE TIME SERIES CLASSIFICATION PROBLEMS

LIM YEE LYN

PROJECT SUBMITTED IN PARTIAL FULFILMENT FOR THE DEGREE OF MASTERS OF DATA SCIENCE

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2024

RANGKAIAN MELEDINGKAN MASA DINAMIK DAN DUA PERHATIAN UNTUK MASALAH KLASIFIKASI SIRI MASA PELBAGAI

LIM YEE LYN

PROJEK YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN
DARIPADA SYARAT MEMPEROLEH IJAZAH
SARJANA SAINS DATA

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI
2024

DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

22 July 2024

LIM YEE LYN
P128731

ACKNOWLEDGEMENT

Undertaking this Master is a great challenge and opportunity. It would not have been possible without the support and help that I have received.

I wish to express my most sincere gratitude and appreciation to Professor Dr. Salwani Abdullah for her patience in providing me with knowledge, invaluable advice, and assistance throughout this dissertation. I am extremely fortunate to have her as my supervisor, as with her guidance, I can complete this dissertation.

A big thank you to all my professors, lecturers, and postgraduate coursemates at UKM for their constant guidance and sharing of knowledge throughout the course.

I am very grateful for sponsorship from the organisation where I work, KPJ Tawakkal Health Centre. Also, deep appreciation to my employer and colleagues for their assistance throughout this period.

Last but not least, thank you to my family who provided continuous support to enable me to complete this dissertation.

ABSTRAK

Algoritma sedia ada untuk pengelasan siri masa digunakan kebanyakannya pada data siri masa univariat. Banyak masalah berkaitan situasi kehidupan sebenar melibatkan siri masa multivariate. Terdapat kekurangan algoritma dengan prestasi yang konsisten dalam MTSC. Satu cara untuk menambah baik MTSC ialah dengan memperluaskan algoritma UTSC yang menjanjikan kepada MTSC. Dynamic time warping (DTW) ialah kaedah yang mantap untuk TSC dan masih disyorkan sebagai penanda aras walaupun terdapat banyak algoritma yang lebih baharu. Salah satu algoritma terkini dengan prestasi baik pada MTSC ialah Rangkaian Perhatian Ganda (DA-Net). Dalam disertasi ini, kami menyiasat pelanjutan kaedah sedia ada DTW ke dalam algoritma baharu, DA-Net. Kaedah yang dicadangkan ini dibandingkan dengan varian DTW, ShapeDTW-1NN untuk kebolehgunaan dalam MTSC. Kedua-dua kaedah boleh digunakan pada MTSC. Perbandingan ShapeDTW-1NN dan DTW-DA-Net menunjukkan bahawa menggabungkan DTW kepada DA-Net yang merupakan algoritma yang lebih terkini mempunyai hasil yang lebih baik daripada varian DTW dengan klasifikasi tradisional, ShapeDTW-1-NN. Penggabungan DTW ke dalam DA-Net meningkatkan prestasi klasifikasi untuk Pengiktirafan Aktiviti Manusia dan data audio. Perbandingan DTW-DA-Net dengan DA-Net dan algoritma berasaskan jarak sedia ada menghasilkan dua prestasi terbaik daripada lapan set data dengan ketepatan tertinggi 0.917 pada set data NATOPS dan 0.595 pada set data Arah Gerakan Tangan.

ABSTRACT

Existing algorithms for time series classification are applied mostly to univariate time series data. Many real-life situations related problems involve multivariate time series. There is a lack of algorithms with consistent performance in MTSC. One way to improve MTSC is by extending promising UTSC algorithms to MTSC. Dynamic time warping (DTW) is a well-established method for TSC and is still recommended as a benchmark even though there are many newer algorithms. One of the recent algorithms with good performance on MTSC is the Dual Attention Network (DA-Net). In this dissertation, we investigate the extension of the existing established method DTW into a new algorithm, DA-Net. This proposed method is compared to a variant DTW, ShapeDTW-1-NN for the applicability in MTSC. Both methods are applicable to MTSC. Comparison of ShapeDTW-1NN and DTW-DA-Net showed that combining DTW to DA-Net which is a more recent algorithm has better outcome than the variant DTW with traditional classification, ShapeDTW-1-NN. The incorporation of DTW into DA-Net improved the performance of classification for Human Activity Recognition and audio data. Comparison of DTW-DA-Net with existing distance-based algorithms produced two best performances from eight datasets with the highest accuracy of 0.917 on the NATOPS dataset and 0.595 on the Hand Movement Direction dataset.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

LIST OF ABBREVIATIONS

AF	Atrial Fibrillation
ALS	Amyotrophic lateral sclerosis
AWR	Articulary Word Recognition
BCI	Brain Computer Interfaces
BM	Basic Motion
CNN	Convolutional neural network
DTW	Dynamic Time Warping
DTW-1NN-D	Dynamic Time Warping - 1 Nearest Neighbour
ECG	Electrocardiograms
ED-1NN	Euclidean Distance – 1 Nearest Neighbour
EEG	Electroencephalograms
HIVE-COTE	The Hierarchical Vote Collective of Transformation-based Ensembles for Time Series Classification
HMD	Hand Movement Direction
HB	Heartbeat
MEG	Magnetoencephalography
MTSC	Multivariate Time Series Classification
NATOPS	Naval Air Training and Operating Procedures Standardization
Norm	normalised
ResNet	Residual Networks
RNN	Recurrent Neural Network
ROCKET	Random Convolutional Kernel Transform
SA_DTW	Shape and Angle Dynamic Time Warping
SEWA	Squeeze-Excitation Window Attention
SRS2	Self Regulation SCP2
SSAW	Sparse Self-Attention within Windows
SWJ	Stand Walk Jump
TSC	Time Series Classification

UCR	University of California. Riverside
UEA	University of East Anglia
UTSC	Univariate Time Series Classification
UKM	Universiti Kebangsaan Malaysia
WEASEL	Word Extraction for time series classification
1-NN	1 nearest neighbour classifier

CHAPTER I

INTRODUCTION

1.1 BACKGROUND

Data points collected in sequence in regular time intervals are known as time series data (Liang & Wang 2021). TSC is the fitting of these data points to a discrete response variable (Middlehurst et al. 2023). Classification of time series data involves pattern detection and feature selection and has been very popular in the past few decades (Liang & Wang 2021). Time series classification (TSC) involving only one variable is known as Univariate Time Series Classification (UTSC) whereas if there is more than one variable, it is known as Multivariate Time Series Classification (MTSC) (Chen et al. 2022). The order of these data is not necessarily in time; audio and images can also be transformed to frequency or series for TSC (Middlehurst et al. 2023).

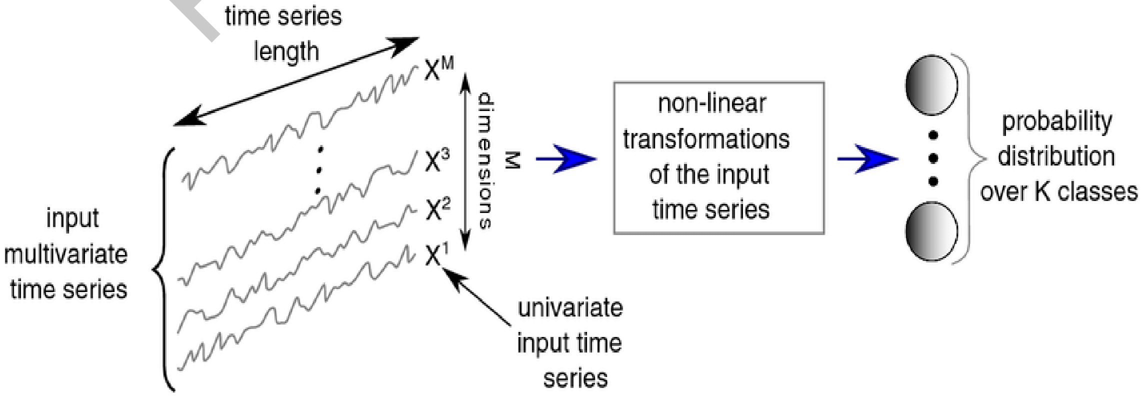

time series length
X^M
X^3
X^2
X^1
M dimensions
input multivariate time series
univariate input time series
non-linear transformations of the input time series
probability distribution over K classes

Figure 1.1 Example of MTSC using deep learning

Source: Ismail Fawaz et al. (2019)

MTSC depends on the extracted features which can be local, global, or spatial. Figure 1.2 shows local features which are temporal features, global features which are the relationship between the local features, and the spatial features which can be calculated such as distance or correlation (Du et al. 2023; Xiao et al. 2021).

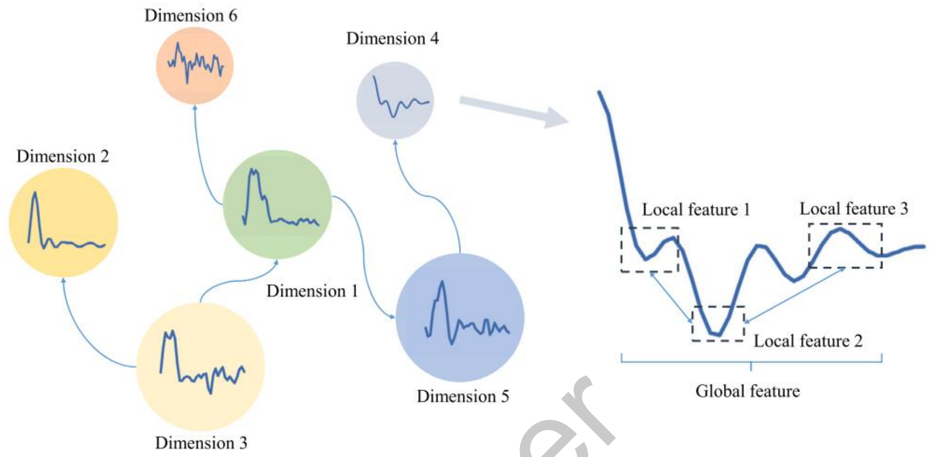


Figure 1.2 Local, global and spatial dependency features. Arrows indicate spatial dependencies.

Source: Du et al. (2023)

With the advancement of technology, the development of wearable devices allows convenient monitoring of parameters such as blood pressure, pulse rate, blood oxygen saturation, and gesture (Qiu et al. 2022). There are many other real-life applications of MTSC. Distance-based methods with k-nearest neighbour (1-NN) demonstrated that it is the best method for small datasets (Deng et al. 2013).

Many MTSC performed in previous years can be distinguished as traditional or deep-learning based and by methods such as distance-based (Euclidian Distance (ED) and Dynamic Time Warping (DTW)), shapelet-based, dictionary-based (Boss, WEASEL + MUSE) and interval-based methods (Chen et al. 2022).

The latest taxonomy presented by Middlehurst et al (2023) classifies TSC into eight categories (as summarised in Figure 1.3); distance based, feature-based, interval-based, shapelet-based, dictionary-based, convolution-based, deep learning-based, and hybrid approaches (Middlehurst et al. 2023). Distance-based measures distance between two series. Feature-based extracts global features and passes to the classifier. Interval-based selects phase-dependent intervals and converts them to features.

Shapelet-based produces phase-independent sub-series. Dictionary-based counts repeating patterns as a feature for the classifier. Convolution-based create feature patterns by convolutions and pooling. Deep learning is based on neural networks, whereas hybrid is the combination of two or more approaches (Middlehurst et al. 2023).

One concern on MTSC is whether to normalize the dataset or not. In general cases, normalization helps to standardize comparison between datasets. However, for time series data, there is a complex relation between the points and shape where normalization may affect classification (Ruiz et al. 2021). ROCKET, TapNet, InceptionTime, and ResNet have internal normalization. The comparison of results may not be fair to the baseline DTW. A comparison of normalizing and not normalizing data by Ruiz et al showed that DTW performed worse on normalization and HIVE-COTE showed no significant difference. Therefore, the authors conclude that normalization is not necessary and will not create bias in the baseline DTW classifiers (Ruiz et al. 2021).

Elastic similarity and distance measures can compute similar series together by realigning the temporal misalignments which are mainly applicable to univariate time series datasets. Shifaz et al. (2023) used seven common elastic similarity and distance measures showed that the use of independent and dependent DTW calculation affects the outcome where the result is consistent. A certain type of dataset will perform better with a certain type of calculation. DTW-dependent performs poorly when there are many dimensions in the dataset.

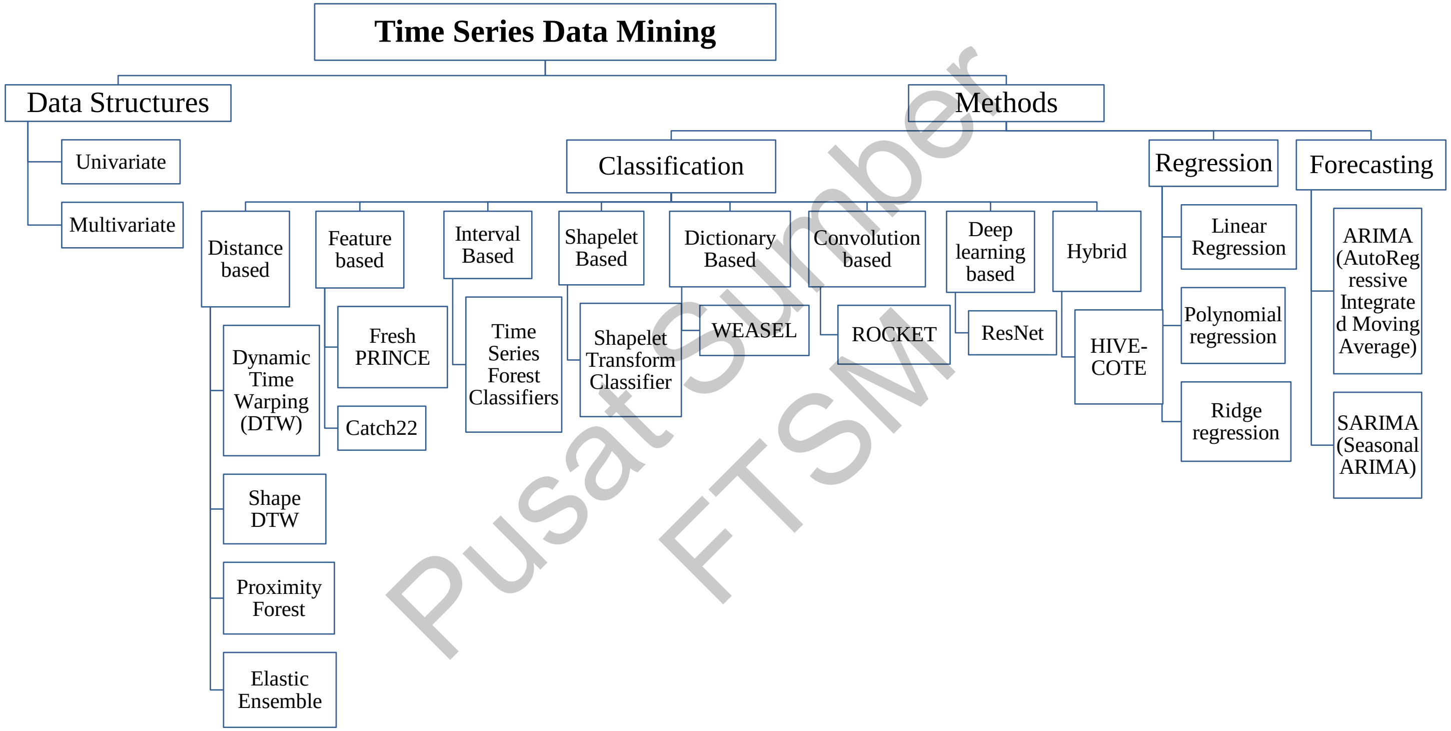


Figure 1.3 Taxonomy of Time Series Data Mining

The recent algorithms with improvement are mainly focused on UTSC. MTSC differs from UTSC as the different dimensions may have interaction. There may be more MTSC problems in real-life scenarios such as electrocardiogram interpretation and human activity recognition. Ruiz et al. (2021) suggest that the simplest approach to overcome the difficulty if multivariate data cannot be processed by an algorithm is to adapt a univariate classifier to the MTSC. In 2018, a standard UEA archive for 30 MTSC problems was made available (Bagnall et al. 2018). Ruiz et al (2021) performed simple adaptations of UTSC approaches on 26 of the equal-length MTSC problems in the UEA archive. The outcome showed that it was difficult to overcome the results of DTW. Four algorithms are better than DTW but it was a comparison of only 26 datasets and may not be generalized to other cases. The benchmark distance-based approach in TSC is DTW, but there are also datasets where this is not true. However, it is commonly used for baseline as a reference and recommended as an initial benchmark for MTSC (Ruiz et al. 2021; Shokoohi-Yekta et al. 2017). Figure 1.4 shows how two time series are aligned with minimal distance using DTW.

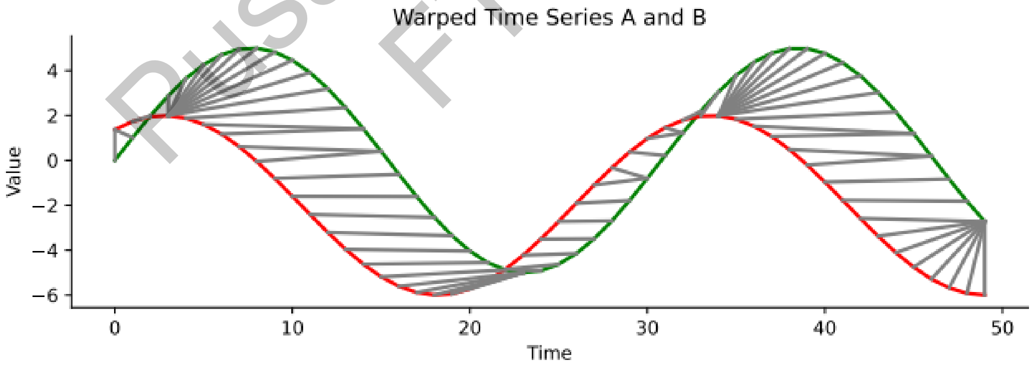


Figure 1.4 Example of how DTW compensate shape shift by realigning two series.

Source: Middlehurst et al. (2023)

1.2 RESEARCH PROBLEM

There are many distance-based methods performed on UTSC that perform well, but not established in MTSC. Many of the distance-based methods such as FastDTW (Salvador & Chan 2007), ShapeDTW-1NN (Zhao & Itti 2017) and Proximity Forest (Lucas et al. 2019) are performed on UTSC. For MTSC, methods performed are such as ED-1NN, DTW-1NN, and their variants (Chen et al. 2013). We would like to investigate the applicability of UTSC methods on MTSC.

Distance measurement between time series to assess similarities can be performed by Euclidean distance or DTW. There is a limitation in Euclidean distance for time series data and DTW is more suitable. DTW is a measurement of distance between time series data and can be used to detect similarities in time series.

Dau et al. (2018) concluded that the amount of warping determines the performance of the data mining problem and the best warping window is non-transferable between any two tasks (such as for clustering and classification). The data structure, time series shape, and dataset size determine the optimal warping window or value for warping. For supervised cases with sufficient data, cross validation with warping during training is recommended, but is not recommended for small dataset and unsupervised cases. The common method to select a warping window for TSC is via cross-validation over increments of the warping window as it is simple, parameter-free, and works well in practice. However, in some cases, this may not be the best way to determine the window. It is suggested that strengthening DTW as a baseline can further improve the outcome of more sophisticated methods, especially in real-world problems where minor improvements in accuracy matters (Dau et al. 2018).Therefore, we would like to investigate the effect of utilizing DTW in recent algorithms to see if there will be a difference in performance.

The difficulties in MTSC have been challenged by many existing methods. DA-Net is one of the recent methods that produced promising outcomes, with a limitation of fixed window size that may not be applicable to all real-life datasets. DTW analysis may be beneficial as an initial step before adjusting window size.

1.3 RESEARCH QUESTIONS

1. Are distance-based methods such as DTW and shapeDTW performed on univariate time series analysis applicable to multivariate time series?

2. Does the incorporation of DTW to DA-Net affect the performance of multivariate time series classification?

3. How is the performance of DTW-DA-Net compared with other state-of-the-art multivariate time series classification methods?

1.4 RESEARCH OBJECTIVES

1. To assess the applicability of univariate distance-based classification methods in solving multivariate time series classification.

2. To propose an incorporation of DTW with DA-Net; and compare to DA-Net.

3. To evaluate and compare the performance of DTW-DA-Net with other state-of-the-art multivariate time series classification methods.

1.5 DISSERTATION ORGANIZATION

This dissertation is divided into five chapters. Chapter I consists of the background, research problems, research questions and research objectives.

Chapter II explores the problems in TSC, both in univariate and multivariate time series problems. This chapter also studies the current approaches for TSC problems with an in-depth exploration on distance based, DTW, and attention-based, DA-Net algorithms.

Chapter III contains the research design, datasets, methods, and evaluation. The research design gives an outline of the proposed algorithms and evaluation used to test the objectives. The datasets are categorized by data type. DTW method, DA-Net algorithm, and parameters used for evaluation are explained.

Chapter IV shows the result and discusses the findings with relation to the research questions and objectives.

Chapter V concludes the dissertation with the research summary, limitations, and future work.

CHAPTER II

LITERATURE REVIEW

2.1 INTRODUCTION

This chapter of eight sections reviews available literature for TSC; the problems, and the algorithms performed. The reviews focus on classification problems, UTSC, MTSC, DTW, and DA-Net.

Classification problems are introduced in section 2.2. Sections 2.3 and 2.4 discuss the different approaches for UTSC of MTSC respectively. Section 2.5 studies the DTW approach. Self-attention and transformer-based methods are discussed in Section 2.6. Discussion in Section 2.7 shares the problems in literature and methods to overcome them. The summary in Section 2.8 concludes the literature review.

2.2 CLASSIFICATION PROBLEMS

TSC involves the use of machine learning algorithms to correctly predict and classify data from real values to the class labels. There has been increased interest in time series analysis in recent years. Traditional methods such as Random Forest and Support Vector Machines do not consider time series points in sequence when input into the method. Therefore, these methods are not suitable for TSC (Hao et al. 2021).

Initial effective TSC methods are distance-based methods. Distance-based methods search for obvious patterns in the time series and classify them using 1-NN. The initial distance measure uses Euclidean distance. Minor changes in the series will affect the alignment causing inaccurate results. DTW is the choice for TSC due to its effectiveness in aligning time series.

Multi-layer perceptron (MLP) is an architecture of fully connected neurons with layers and neurons as hyperparameters. It works by a feedforward mechanism that processes input in a fixed manner. This results in the limitation of MLP in TSC where it is unable to detect temporal dependencies in time series data (Mohammadi Foumani et al. 2024). To overcome this, Iwana et al. (2020) used feature extractor, DTW which has an elastic matching ability. Compared to the dot product within a neuron that has stagnant weight mapping, DTW allows weights to be aligned into the layers. This method proved beneficial in overcoming the feedforward mechanism, allowing temporal patterns and variable length recognition in time series (Iwana et al. 2020). Apart from limitations on temporal features, MLP also cannot capture multi-scale input, trends, and fluctuations and therefore unable to handle irregular time series. For irregular data, CNN and transformers are more suitable (Mohammadi Foumani et al. 2024)

CNN networks have improved in recent years due to the increase in depth of the newer algorithms (Gu et al. 2018). Other means of improvement include using smaller convolutional filters, the addition of pooling layers for feature map dimension reduction, and batch normalization (Mohammadi Foumani et al. 2024). This increase allows better feature representation but also increases the risk of overfitting and increases difficulty in optimization. (Gu et al. 2018). Another disadvantage of CNN networks is that they use local data to produce features and therefore unable to capture long-term dependencies (Hao & Cao 2020).

Recurrent neural network (RNN) is built with internal memory with a feed-forward network. CNN is suitable for detecting spatial relationships which enables it to capture correlations of channels in the series whereas RNN captures temporal dependencies from previous data to predict upcoming values (Mohammadi Foumani et al. 2024). The advantages of both CNN and RNN make it a good hybrid method if used together as they overcome each other’s disadvantages. Combining the methods allows both temporal and spatial features to be captured.

The hybrid method of RNN and CNN can detect local and temporal correlation but is unable to detect the long-range dependencies of the time series (Mohammadi

Foumani et al. 2024). In time series data, current points may be influenced by the previous points and can influence future points. The short-term and long-term dependencies is relative to when the points are affected (Hao & Cao 2020). It is possible for deep CNN and RNN to capture long historical data but is not justifiable due to the high computational requirements and there are still problems in capturing the data (Hao & Cao 2020). Another problem in MTSC using CNN and RNN is that the relation between the multiple variables does not appear in the feature extraction.

Attention-based models have a broader field and can capture long-range dependencies (Hao & Cao 2020; Mohammadi Foumani et al. 2024). This method works by only extracting relevant features and ignoring irrelevant data.

2.3 UNIVARIATE TIME SERIES CLASSIFICATION

Bagnall et al. (2017) published a review in 2017 to compare 18 TSC algorithms on 85 UCR datasets. This is then known by many as the 'bake off' for TSC. It categorized algorithms by feature extracted from time series data to whole series, distance, intervals, shapelets, dictionary, convolutional, combinations and model-based (Bagnall et al. 2017.; Shifaz et al. 2023). From then, the UCR dataset has increased to more than 100 datasets and there are many new algorithms. Table 2.1 summarizes some of the algorithm types performed in recent years.

Distance-based methods for UTSC are usually used with 1-NN. Apart from DTW-1NN, improved distance similarity approaches are Elastic Ensemble (EE) (Lines & Bagnall 2015) and Proximity Forest (PF) (Lucas et al. 2019). Current distance-based methods can realign similar series using elastic adjustment.

Table 2.1 Summary of literature review on algorithm types in UTSC.

Algorithm		**Algorithm Type**
WEASEL v2.0	WEASEL2.0 with Dilation(Schäfer & Leser 2023)	Dictionary-based
FreshPRINCE	Fresh Pipeline with Rotation Forest Classifier (Middlehurst & Bagnall 2022)	Feature-based
HC2	HIVE-COTE version 2(Middlehurst et al. 2021)	Hybrid
ROCKET	Random Convolutional Kernel Transform(Dempster et al. 2020)	Convolution based
ShapeDTW	Shape Based DTW(Zhao & Itti, 2019)	Distance-based
EE	Fast Elastic Ensemble (Oastler & Lines 2019)	Distance-based
PF	Proximity Forest (Lucas et al. 2019a)	Distance-based
Catch22	Canonical Time Series Classification (Lubba et al. 2019)	Feature-based
WEASEL	Word Extraction for TSC (Schäfer & Leser 2017a)	Dictionary-based
ResNet	Residual Network(Wang et al. 2017)	Deep Learning based
STC	Binary Shapelet Transform Classifier(Bostrom & Bagnall 2017)	Shapelet-based
TSF	Time Series Forest(Deng et al. 2013)	Interval-based

DTW-1NN has limited information on the temporal features that segregate different classes of time series. TSF is an interval-based method that can detect the temporal patterns and distortion of the axis. It is based on entrance (entropy and distance) gain. The limitation of this method is the computational complexity when the feature space is large (Deng et al. 2013).

The dictionary-based method uses the bag-of-word concept. WEASEL (Schäfer & Leser 2017a) is an example of this type of method. Current TSC difficulties include inability to produce high accuracy with large data, methods which are scalable but compromised accuracy, or models with accuracy that is not scalable to large data. WEASEL extracts feature vectors and performs sliding windows for TSC. It is scalable and accurate. There is a disadvantage where it leaves a huge memory footprint and therefore limits its uses. This leads to the improvement of the model to WEASEL 2.0 (Schäfer & Leser 2023) that incorporates dilation and hyperparameter ensembling resulting in improved performance with a fixed memory footprint.

Random Convolutional Kernel Transform (ROCKET) is a kernel-based method that performs convolution to transform data for classification with logistic regression.

This method highlights the ability to train large datasets with minimal time (Dempster et al. 2020)

Deep learning methods can be RNN or temporal convolution. ResNet and InceptionTime are examples of temporal convolution-based methods (Shifaz et al. 2023).

2.4 MULTIVARIATE TIME SERIES CLASSIFICATION

The initial compilation of the MTSC dataset was by Baydogan[1] but the data are small, have variable length, are dependent, and not a good representation for MTSC. From year 2018, 30 UEA MTSC by Bagnall et al (2017) allows focusing on classification tasks instead of preprocessing. Twenty-six of these datasets are of equal length.

There are many MTSC performed in previous years using distance-based methods and nearest neighbour classification. In traditional methods, the same distance metric is used for datasets of various varieties which may result in lesser accuracy (Chen et al. 2021). Newer methods focus on time and less on the frequency domain. Another limitation of traditional practice is the assumption of a linear relationship on multivariate time series data. To overcome the limitation, Chen et al. (2021) performed time-frequency deep metric learning (TFDM) model which can learn nonlinear and meaningful distance of MTS data and develop multilevel time-frequency representation resulting in 18 public MTS datasets performing better than other state-of-the-art methods (Chen et al. 2021).

Karim et al. (2018) extended and augmented the existing UTSC classification methods to MTSC, namely Long Short-Term Memory Fully Convolutional Network (LSTM-FCN) and Attention LSTM-FCN (ALSTM-FCN). (Karim et al. 2018) The models produced better results than many of the existing models with the additional benefit of minimal preprocessing (Karim et al. 2018).

[1] http://www.mustafabaydogan.com/multivariate-time-series-discretization-for-classification.html

Table 2.2 Summary of literature review on approaches for MTSC.

Work	**Author**
Alignment-driven Neural Network	(Bignoumba et al. 2024)
Multi-feature based network	(Du et al. 2023)
Dual-attention network	(Chen, R et al. 2022)
Time-frequency deep metric learning	(Chen, Z et al. 2021)
TapNet	(Zhang et al. 2020)
Long Short-Term Memory Fully Convolutional Network	(Karim et al. 2018)
Time Series Forecast	(Deng et al. 2013)
Discrete SVM and fixed cardinality warping	(Orsenigo & Vercellis 2010)

Orsenigo & Vercellis, (2010) proposed to use a combination of fixed cardinal warping distance with SVM to perform on MTSC of variable length. Firstly, the time series are converted to the same length by fixed cardinality of warping distance where the output is rectangularised to match similar pairs of time series. Then, the processed information is fed to a discrete SVM for regularization and classification. Four benchmark datasets and two real-world marketing data were used to produce increased accuracy of 0.5% to 3.1% compared to traditional SVM and 1-NN (Orsenigo & Vercellis 2010).

Tapnet and LSTM-FCN are examples of deep learning method (Chen et al. 2022). Even though deep learning methods can directly map low-dimension features to high-dimension features from raw data and there is a benefit of the lower requirement of a domain expert, there is a limitation in managing long sequences which is common in real-life situations (Chen et al. 2022) .

There can be irregularities or inconsistencies in univariate and multivariate time series data due to the inability to obtain complete raw data which can cause sparse data and reduce the performance of traditional and deep learning models performance (Bignoumba et al. 2024). This is usually resolved by imputing or interpolating data but with the disadvantage of introducing noise into the data. Bignoumba et al. (2024) introduced a new deep neural network model called Alignment-driven Neural Network (ALNN) which uses a duplication process and exponential decay mechanism to convert irregular multivariate time series data into a pseudo-aligned (or pseudo-regular) latent values before passing the data to RNN model. RNN is the recommended model for regular time series data (Bignoumba et al. 2024; Bińkowski et al. 2017) ALNN produced satisfactory results on MIMIC-III healthcare data in predicting mortality. The aggregation method in an hourlong bin to overcome the missing or inconsistent data problem can result in loss of data or fine-grained information (Bignoumba et al. 2024) The assumption in basic RNN models is that data is evenly spaced out, but this is not the case is real life. Bayesian Network (MacKay 1992), Gaussian Processes (Roberts et al. 2013), and Support Vector Regression (Vapnik & Golowich, 1996.) failed to overcome this as the models cannot capture complex temporal dependencies (Bignoumba et al. 2024).

A time series attentional prototype network (TapNet) combines the traditional and deep learning methods. It can manage limited training labels and extract low-dimensional features. The approach shows good outcomes on 18 UEA Multivariate time series compared to eight state-of-the-art baseline methods (Zhang et al. 2020).

Traditional learning such as Time Series Forecast (TSF) which is a tree-ensembled method that uses simple features such as mean, standard deviation and slope showed good performance with the benefit of computational efficiency compared to a one-nearest neighbor classifier with DTW (Deng et al. 2013; Chen et al. 2022). Traditional learning methods such as Bag-of-Patterns and Time Series Shapelet work well with small datasets but have difficulties in handling large multivariate data (Zhang et al. 2020).

Many MTSC gives attention to local and global features, compared to spatial features. Multi-feature-based network which has a global-local block, and a spatial-local block, was able to capture all these effects together and performed well on UEA datasets. The method works by capturing the spatial dependency features with a spatial-local block while integrating the local and global features (Du et al. 2023).

2.5 DYNAMIC TIME WARPING

DTW was introduced as early as 1957 by Bellman (1957) to solve optimization problems using global optimum. Then, in 1978, Sakoe & Chiba used DTW for speech recognition analysis.

Table 2.3 Summary of literature review on DTW-related methods

Work	Author
Python package: dtwParallel	(Escudero-Arnanz et al. 2023)
Parameterization of the cost function of DTW	(Herrmann et al. 2023)
Wavelet-DTW Hybrid Attention Network (WHEN)	(Wang et al. 2023)
Amercing DTW	(Herrmann & Webb 2023)
MTSC bake-off evaluation	(Ruiz et al. 2021)
Generalizing DTW from UTSC to MTSC	(Shokoohi-Yekta et al. 2017)
Shape DTW	(Zhao & Itti 2017)
Shape and inclination angle-based	(Cao & Liu 2016)
Flexible DTW	(Hsu et al. 2015)
Weighted DTW	(Jeong et al. 2011)
FastDTW.	(Salvador & Chan 2007)
Derivative DTW	(Keogh & Pazzani 2001)
DTW in speech recognition	(Sakoe & Chiba 1978)
Dynamic Programming	(Bellman 1957)

DTW itself may cause undesirable alignment and constraints may cause incorrect warping. Keogh & Pazzani (2001) proposed the derivative DTW to overcome these problems by using the properties of the shape for alignment instead of data points.

Fast DTW which was introduced by Salvador & Chan (2007) resulted in a refined projected solution from a course resolution. This method produced better accuracy compared to Sakoe-Chuba Bands. Fast DTW is a form of DTW that has linear time and space complexity. The benefit of Fast DTW is that it can run on much larger datasets than DTW and warp both similar and dissimilar time series.

Weighted DTW is a penalty-based method. The modified logistic weight function is assigned to points with a high difference to minimize the influence of outliers

(Jeong et al. 2011). This method is added to derivative DTW and compared with other approaches using the UCR dataset. There is improved accuracy in both classification and clustering problems using this method.

Hsu et al. (2015) introduced flexible DTW which uses the flexible longest common subsequence where the score is added to a continued one-to-one match of long fragments. The writers used the voting scheme to compare their method with DTW and derivative DTW. Results show that there is a lower average error rate for flexible DTW.

Cao & Liu (2016) proposed a method, SA_DTW, which analyzes the shape feature and tilt angle as features to search for the similarities of DTW in multivariate time series. This method improved the accuracy in matching similarities with the advantage of reduced time.

DTW has high flexibility which can be an advantage or disadvantage in TSC. In cases where it has disadvantages, constraints can be applied to limit this, such as window constraint and weights that limits diagonal points to the cost of the alignment (Herrmann & Webb 2023). However, these two methods have drawbacks. Windows does not allow flexibility beyond the window, allows unconstrained flexibility within the window, and stops abruptly. Weights cause large warping at little cost. Herrmann & Webb (2023) introduced amerced DTW which uses an additive weight that can be tuned to overcome the mentioned disadvantages.

Wang et al. (2023) focuses on the heterogeneity of time series where there is intra-sequence nonstationary and inter-sequence asynchronism. The writers explained that these factors are commonly overlooked. They proposed a hybrid attention network, WHEN which is made out of two modules: WaveAtt and DTWAtt. WaveAtt analyses the nonstationary time series whereas DTWAtt synchronizes the input sequences with a universal parameter. Experiment on the 30 UEA datasets showed that this method performed well.

Escudero-Arnanz et al. (2023) created an open software named dtwParallel for the evaluation of DTW in univariate and multivariate time series data. It considers the parallelization, multiple features, low floor, and high-ceiling design to allow ease of usage for different expertise levels.

1-NN is one of the most used classification methods for TSC; and the distance-based function commonly used together is DTW (Ruiz et al. 2021). An experimental comparison of recent MTSC by Ruiz et al (2021) concludes DTW is still the standard for TSC benchmark as it still performs well even if compared to newer algorithms (Ruiz et al. 2021). DTW can be performed in univariate or multivariate but most research are performed on univariate datasets dataset (Zhao & Itti 2017). DTW usage in MTSC is projected from existing DTW used for univariate. There are two ways for DTW in MTSC which are dependent or independent warping. There are mixed opinions on which option to select. Some opinions suggest that it is not necessary to select the type of DTW as both ways are equivalent. Some research did not state specifically how DTW is generalized to the multidimensional case. There is also a suggestion that both ways can produce different classification outcomes and proper analysis should be performed to select the suitable method for different cases. In this dissertation, we use the dependent approach as we consider the multivariate time series channels to be dependent. DTW from the sktime library and aeon toolkits are based on dependent DTW. Ruiz et al. (2021) used dependent DTW as their baseline on MTSC comparison.

DTW can be explained as an algorithm search which match point-to point by the similarity of Euclidean distance. This matching is of coordinate value have risk of error as distinct local structures can be matched incorrectly. DTW achieves global optimum but can mistakenly pair two series as it may miss out on local matching Zhao & Itti (2017) proposed an improved version, called shapeDTW that matches similar local structures and avoids pairing of different neighbouring structures. The outcome of their method of NN-shapeDTW for the classification of 84 UCR univariate time series datasets showed that it is better than DTW on 64 datasets and improved the classification accuracy of 18 datasets by more than 10% (Zhao & Itti 2017).

ShapeDTW is a series-to-series transformation where the time series is converted into shape descriptors. The shape descriptors are the subset of the time series with representative or neighboring points. It can be the original sequence, slope, mean value, or wavelet. The original sequence consists of parts of similar points with dissimilar shapes. Time series can be split into multiple intervals Piecewise aggregate approximation, PAA is obtained by using the mean of datapoints in each of the divided series. If the gradient of each interval is used as a descriptor, it is termed slope. Discrete wavelet Transform, DWT uses a wavelet as a coefficient incorporated into the sequence to break down and join back sequences to produce the shape descriptor. Derivative uses DTW for shape representation like how slope is used. It is also possible to compound two or more shape descriptors together. Slope and derivative are not affected by shift on the y axis.

The main difference between DTW and shapeDTW is that DTW measures similarities of two-time series by comparing the Euclidean distance between the time points whereas shapeDTW compares the Euclidean distance between the shape descriptors(Zhao & Itti, 2019). Therefore, DTW finds a global optimum whereas shapeDTW investigates local shape differences.

Comparison of UTSC of DTW and variants on artificially simulated aligned path and real audio signals showed that shapeDTW has lower alignment errors compared to DTW, derivative DTW, and weighted DTW. NN-shapeDTW also outperforms NN-derivativeDTW (Zhao & Itti 2017).

Shokoohi-Yekta et al. (2017) suggests that DTW can be extended from UTSC to MTSC via two strategies, which are dependent or independent DTW. The Independent method calculates as a univariate method and sums up the distances whereas dependent DTW calculates each step as multidimensional and uses Euclidean distances for further calculation. They proposed an adaptive method to select the suitable choice but there is insufficient evidence to support the idea based on the outcome.

2.6 DUAL ATTENTION NETWORK

With the concept of channel features in the CNN layer, the writers Hu et al. (2017) were able to modify the dependencies between channels by using a squeeze-and-excitation

(SE) block. The blocks can regulate the features of the channels in a flexible manner and are piled collectively to be applied throughout datasets. The execution time may be prolonged but the benefit of using SE blocks in state-of-the-art CNN justify this disadvantage.

Table 2.4 Summary of literature review on self-attention and transformer-based methods.

Work	**Author**
Self-attention and relative positioning attention blocks	(Abbasi & Saeedi 2023)
Dual-attention network for MTSC	(Chen et al. 2022)
Multi-branch CNN with Squeeze and Excitation Attention Blocks	(Altuwaijri et al. 2022)
Transformer encoder for multivariate time series learning	(Zerveas et al. 2021)
Swin Transformer	(Liu et al. 2021)
CTNet	(Lian et al. 2021)
MSCRED	(Zhang et al. 2019)
Squeeze and Excitation Networks	(Hu et al. 2017)

Research has proposed that the modelling of local-global features is of importance. Multi-Scale Convolutional Recurrent Encoder-Decoder (MSCRED) is an algorithm to detect and diagnose anomaly in time series data. It is made of a dual network to create a feature map that is fed into a convolutional decoder. The testing on an artificial dataset and real dataset concluded that the modelling of local and global features works better than state-of-the-art baseline methods (Zhang et al. 2019).

Conversational transformer network, CTNet is used in modelling intra-modal and cross-modal interactions. It uses a transformer-based structure for conversation emotion recognition and produced a 2.1 to 6.2% increase in performance compared to F1-score of state-of-the-art methods (Lian et al. 2021). This showed that the global and local-features modelling is beneficial.

Squeeze and excitation attention blocks is used together with multiple branch convolutional neural network in electroencephalography motor imagery classification. It was tested on two public datasets and performed 10% better than the baseline method in accuracy comparison. This is due to the self-attention blocks and selective reduction of hyperparameters of the multiple branches. Even though the proposed model has 3.9 times more hyperparameters compared to the baseline, result generation is fast due to the parallel processing (Altuwaijri et al. 2022).

Abbasi & Saeedi (2023) proposed to improve TSC by introducing two attention-based processing blocks and claimed that it applies to any Deep Neural Network (DNN) TSC. First block is Global Temporal Attention, and another is Temporal Pseudo-gaussian augmented Self-attention. These blocks are incorporated into Fully Convolutional Network, ResNet, and InceptionTime and tested on UEA datasets resulting in improved accuracy and average rank. The writers highlighted that there is a difference in the range of improvement which is not consistent with the task (Abbasi & Saeedi 2023).

Transformer encoder architecture has been performed by Zerveas et al (2021) on multivariate time series regression and classification. It is the first unsupervised learning method on multivariate time series data which produced better results than fully supervised learning methods (Zerveas et al. 2021).

Liu et al. (2021) proposed Swin Transformer which is a hierarchical Transformer with shifted windows for computer vision. The shifted windows improved efficiency as they limit self-attention computation on non-overlapping local windows. The authors proposed that this method can be useful for all MLP architectures (Liu et al. 2021)

The first unsupervised learning using transformer encoding by Zerveas et al. (2021) and hierarchical, self-attention Swin Transformer by Liu et al. (2021) showed beneficial for MTSC. This has led to the development of a Dual-Attention Network (DA-Net) by Chen et al. (2022) which consists of Squeeze-Excitation Window Attention (SEWA) and Sparse Self-Attention within Windows (SSAW) This method showed improved effectiveness in MTSC when compared to state-of-the-art methods.

Based on the transformer model by Liu et al. (2021) and Zerveas et al. (2021); Chen et al. (2022) proposed to use Swin transformer for MTSC due to its ability to scale down the architecture, able to perform shifted window self-attention of non-overlapping windows and manage the global long-range dependencies of MTSC. However, there are two difficulties with the method where it is unable to detect local discriminating features and there is increase in calculation complexity when the window size increase. The first problem can be overcome by Squeeze-Excitation Window Attention (SEWA) which aggregates feature window by squeezing and then uses MLP for excitation to keep the local key sequences in a feature window of different scale and weightage. To overcome the second problem of Swin Transformer, Sparse Self-Attention within Windows (SSAW) is performed where only activated/higher weighted dot-product scores calculated by Kullback-Leibler and timestamps that do not overlap are selected. This reduces complexity in calculation even if there is an increase in window size. Chen et al. (2022) combined SEWA and SSAW to produce a dual-attention network (DA-Net) for MTSC which showed effectiveness in MTSC compared to state-of-the-art approaches.

Chen et al. (2022) performed further analysis on the effect of window size on PEMS dataset. The dataset was selected randomly. Different window sizes (32,48,64,96) were performed resulting in better results of PEMS which may have DTW important impact. They concluded that the window size can have a non-negligible effect on the outcome.

2.7 DISCUSSION

Traditional methods such as Random Forest and support vector machine are not suitable for TSC as it is unable to capture the sequence in time series data. Initial TSC methods uses Euclidean distance with 1-NN but does not produce promising outcome as minor changes in the series affects the series alignment. DTW is the recommended distance-based method to be used with 1-NN as the standard baseline TSC.

CNN captures spatial dependencies and RNN captures temporal dependencies. Combining both methods produced a hybrid which captures both spatial and temporal features for improved performance. The hybrid method enables algorithms to overcome other method’s limitations.

There are more established and upgraded versions of UTSC algorithms compared to MTSC. It is suggested that algorithms used in UTSC can be extended to MTSC. DTW can be used for UTSC or MTSC. For the extension of DTW from univariate to multivariate data, there are two methods to calculate the distance, dependent DTW and independent DTW. There is no standard recommendation on which DTW calculation to use.

Existing transformer-based methods have limitations on detecting local discriminating features and increase in complexity when there is an increase in window size. Chen et al (2022) proposed DA-Net which consists of SEWA and SSAW to overcome these limitations.

2.8 CHAPTER SUMMARY

This chapter studies the background, history, and limitations of TSC. Reviews of available methods and recent methods are compared. The summary of literature review listed in Tables 2.1, 2.2, 2.3, and 2.4 are presented in Sections 2.3, 2.4, 2.5, and 2.6 respectively. The discussion explains the findings from the literature on the proposed method.

The next chapter explains the research design, dataset, method, and evaluation of the proposed method.

CHAPTER III

MULTIVARIATE TIME SERIES CLASSIFICATION WITH DYNAMIC TIME WARPING AND DUAL ATTENTION NETWORK

3.1 INTRODUCTION

This chapter is made up of six sections that explain the design of this dissertation to address the three research questions; "*Are distance-based methods such as DTW and shapeDTW performed on univariate time series analysis applicable to multivariate time series?*", "*Does incorporation of DTW to DA-Net affect the performance of multivariate time series classification?*", and "*How is the performance of DTW-DA-Net compared with other state-of-the-art multivariate time series classification methods?*".

Section 3.2 explains the research design. The eight datasets used are classified accordingly to four data types and described in Section 3.3. A detailed explanation of the algorithms is in Section 3.4.

Section 3.4 is split into two sections: DTW and DA-Net. Section 3.4.1 on DTW gives examples of numerical calculation. Section 3.4.1 (a), (b), and (c) are definitions from the literature. Section 3.4.1(d) explains the difference in the calculation for multivariate time series data and section 3.4.1(e) shows how DTW for UTSC is modified for MTSC. Section 3.4.1 (f) shows the variant of DTW. The second part of Section 3.4 focuses on DA-Net, which is split into two sections for explanation of the SEWA and SSAW layer.

Section 3.5 defines the performance and evaluation methods used, followed by a chapter summary in Section 3.6.

3.2 RESEARCH DESIGN

We would like to investigate the integration of the distance-based method, DTW into existing algorithms to produce better output. The distance-based method selected was a well-known benchmark used in TSC in previous years. Even with new algorithms in recent years, Ruiz et al. (2021) still propose to use DTW as the initial benchmark for MTSC in their experimental evaluation. Therefore, in this dissertation we used DTW to incorporate into one of the recent algorithms which have produced promising result, DA-Net. Ruiz et al. (2021) stated that one of the top classifiers that works well with DTW is 1-NN. In this dissertation, we selected a variant, ShapeDTW which has produced good results in UTSC to work with 1-NN.

The research design is summarized in Figure 3.1. This dissertation is designed in a way to assess the three objectives proposed. The eight UEA datasets will be used to test the objectives. Details of the methods and algorithms used are in Section 3.4.

The initial step is performing MTSC using the proposed DTW-DA-Net and ShapeDTW to assess the usage of UTSC distance-based methods on MTSC. This will answer the first research question "*Are distance-based methods such as DTW and shapeDTW performed on univariate time series analysis applicable to multivariate time series?*". The percentage difference in accuracy in the methods will be evaluated and explained in chapter four.

Further analysis to investigate the second research question "*Does incorporation of DTW to DA-Net affect performance of multivariate time series classification?*" compares DTW-DA-Net and DA-Net The evaluation metrics are discussed in chapter 3.5 and results presented in chapter 4.3.

The third research question compares DTW-DA-Net and ShapeDTW with existing distance-based algorithms to answer, "*How is the performance of DTW-DA-Net compared with other state-of-the-art multivariate time series classification methods?*". The accuracy is compared using the Wilcoxon signed ranked test and summarized in a critical diagram which is discussed in chapter 4.4.

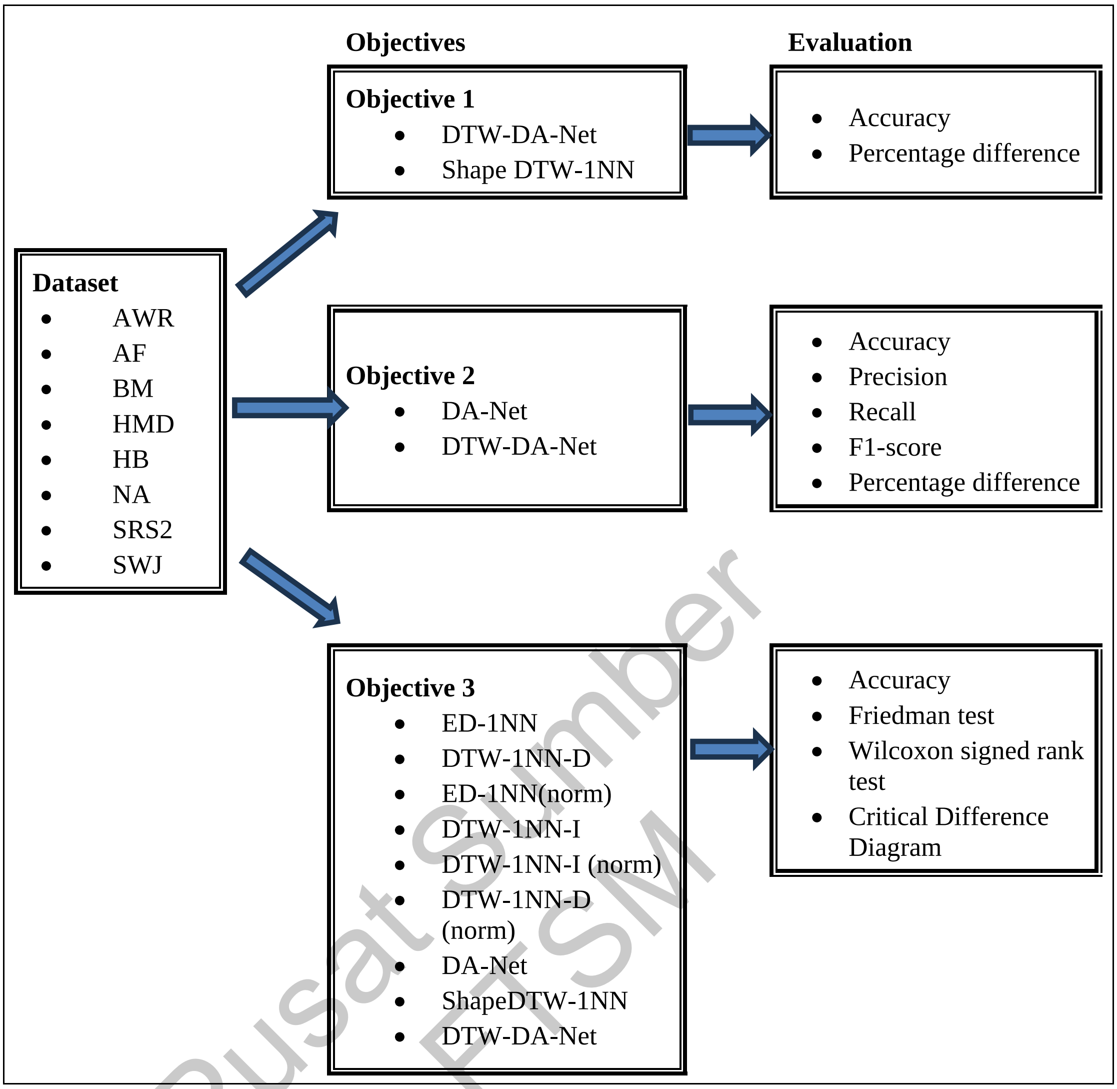


Figure 3.1 Research Design

3.3 DATASET

Benchmark dataset obtained from UEA[1] with a repository of 30 datasets and data size ranging from 27 to 50,000 that covers different areas: human activity recognition (HAR), motion, electrocardiogram (ECG), electroencephalograms/ magnetoencephalography (EEG/MEG), and Audio Spectra (AS) (Bagnall et al. 2024). These data types can be categorized into electrical biosignals, accelerometer/gyroscope generated, coordinates, and audio. Other data types available but not used in this dissertation are such as spectrometry, photometry, and bespoke hardware (Ruiz et al. 2021).

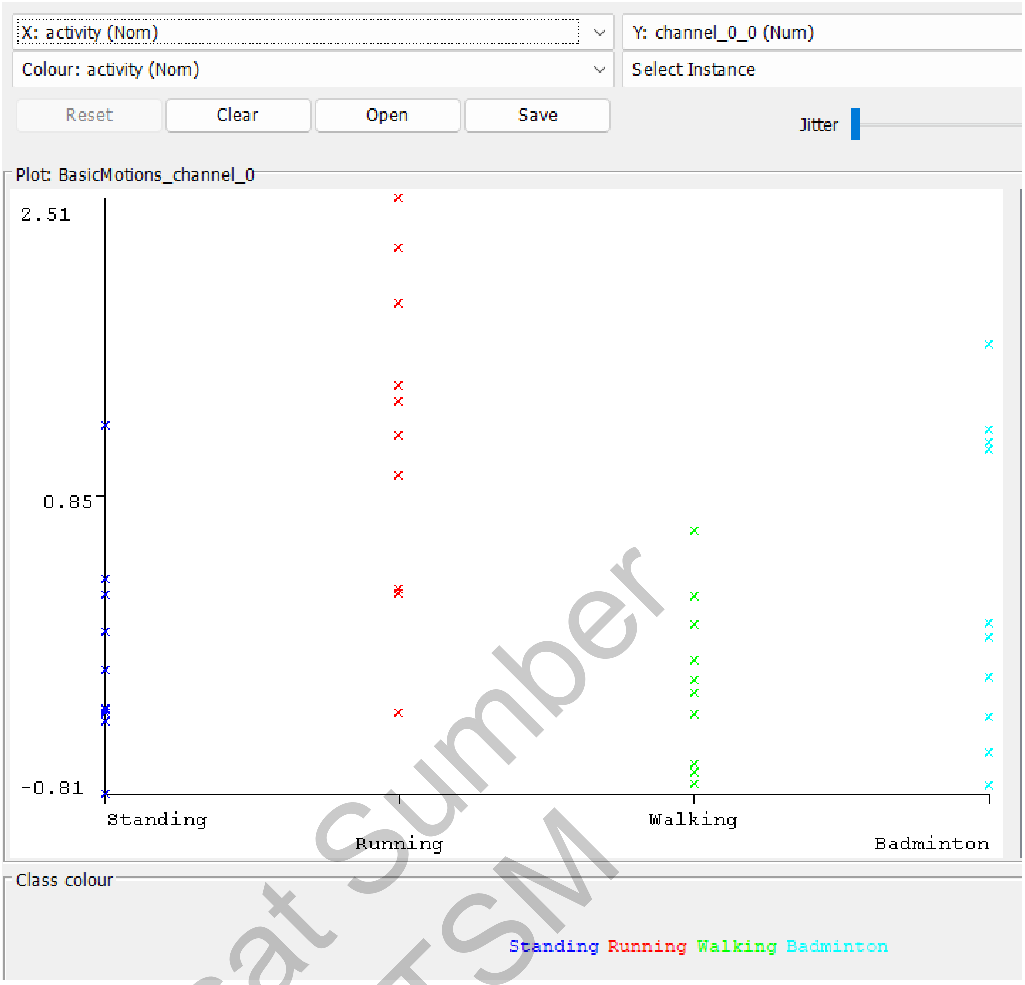
[1] https://www.timeseriesclassification.com/

In this dissertation, we focus on the fixed-length dataset and omit the variable-length dataset. Due to computational limitations, very large dataset is also omitted as they are unable to complete the process in 48 hours. Selected datasets used are shown in Table 3.1.

Table 3.1 Eight publicly available UEA MTSC datasets

Dataset	**Character**					
	Abbreviation	**Train**	**Test**	**Channel**	**Attribute**	**Classes**
Articulary Word Recognition (Normalised)	AWR	275	300	9	144	25
Atrial Fibrillation	AF	15	15	2	640	3
Basic Motion	BM	40	40	6	100	4
Hand Movement Direction (Normalised)	HMD	160	74	10	400	4
Heartbeat	HB	204	205	61	405	2
NATOPS	NA	180	180	24	51	6
Self Regulation SCP2	SRS2	200	180	7	1152	2
Stand Walk Jump	SWJ	12	15	4	2500	3

By referring to the BM dataset in Table 3.1, there are six channels. Each channel has a length of 100 data points representing 100 attributes. There are 40 train samples and 40 test samples for each attribute that are categorized into 4 classes. Figure 3.2 shows an example of the first attribute in channel one where there are a total of 40 samples categorized into the four classes of standing, running, walking, and badminton. Table 3.2 shows the data of the first attribute in the six channels.



X: activity (Nom)

Y: channel_0_0 (Num)

Colour: activity (Nom)

Plot: BasicMotions_channel_0

Figure 3.2 Visualization of channel one, first attribute, with a total of 40 samples in four classes (standing, running, walking, badminton) for BM data.

Table 3.2 Example data of the first attribute from channel one to six

Attribute	Channel 1	Channel 2	Channel 3	Channel 4	Channel 5	Channel 6
1	40 samples	40 samples	40 samples	40 samples	40 samples	40 samples
	Min: -0.814	Min: -3.375	Min: -1.411	Min: -1.159	Min: -0.527	Min: -1.377
	Max: 2.508	Max: 1.783	Max: 1.549	Max: 0.684	Max: 0.743	Max: 0.906
	Mean: 0.323	Mean: -0.156	Mean: -0.185	Mean: -0.055	Mean: 0.044	Mean: 0.127
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
100						

Datasets are categorised according to domain type summarised in Figure 3.3 as mentioned by Ismail Fawaz et al. (2019), Ruiz et al. (2021), and Schäfer & Leser (2017b)

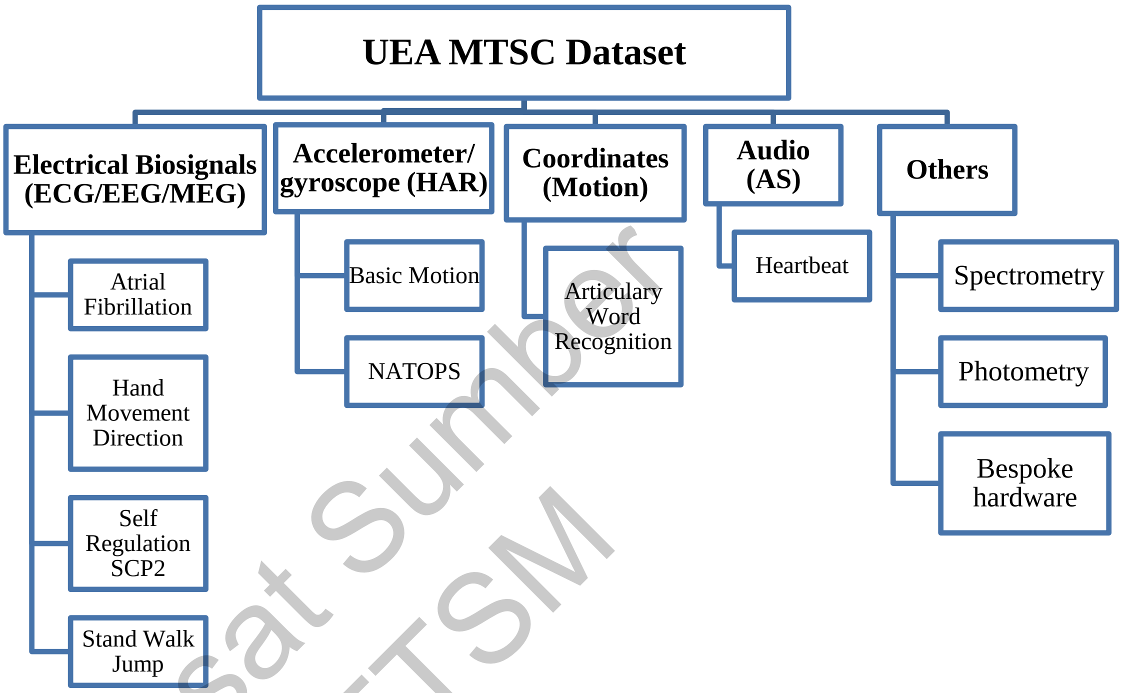


Figure 3.3 UEA Dataset according to data type

Source: Ruiz et al. (2021)

3.3.1 Electrical Biosignals

Many real-life situations use electrical biosignals such as ECG, EEG, and MEG. ECG measures the heart’s electrical activity, EEG measures the brain waves, and MEG measures the brain’s magnetic magnitude. In this dissertation, examples of such datasets are AF, HMD, SRS2, and SWJ (Ruiz et al. 2021).

a. Atrial Fibrillation(Goldberger et al. 2000)

This dataset was created from data used in Computers in Cardiology Challenge 2004 consists of two-channel ECG recordings. It predicts spontaneous termination of AF. Raw instances of 5 s segments of AF, with two ECG signals, each sampled at 128 samples per second; with class labels *n*,*s*, and *t*. Class *n* is a non-termination of AF at least one hour after recording, class *s* is AF that self-terminates at least one minute after recording, and class *t* is immediate termination within 1 s of recording.

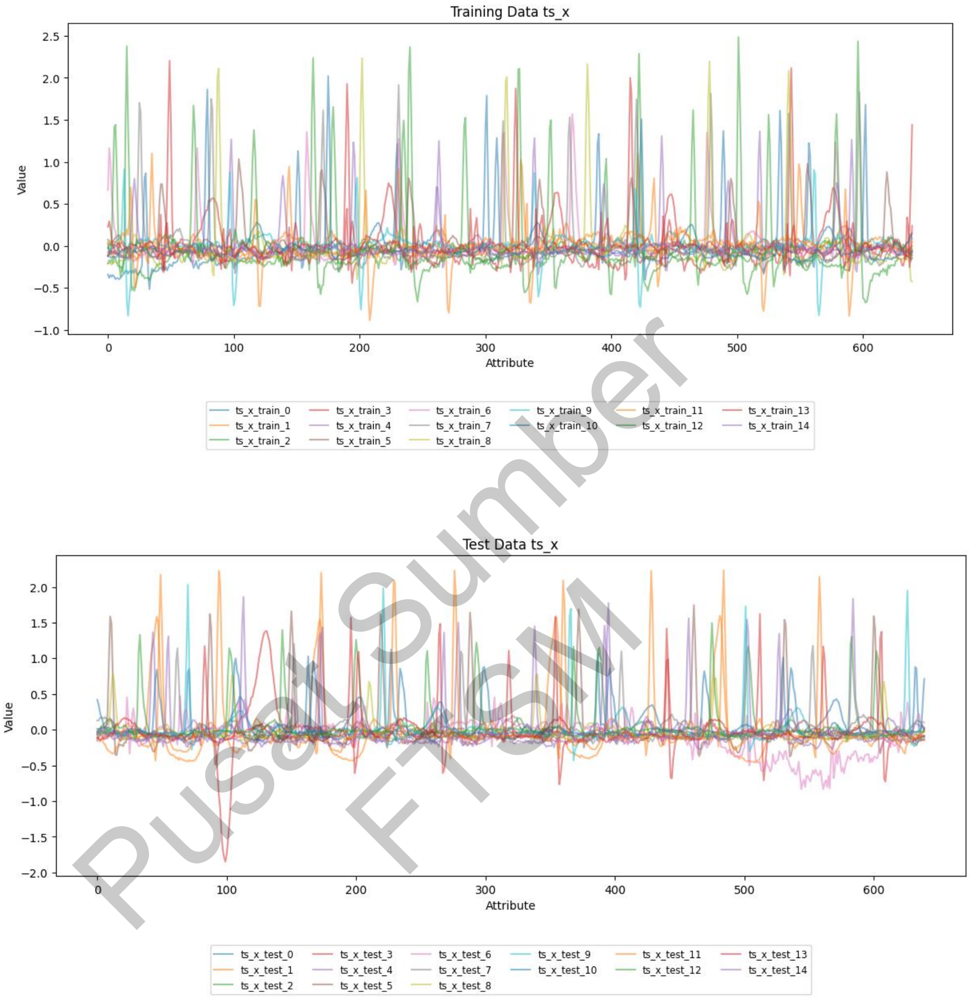


Figure 3.4 Visualization of Atrial Fibrillation dataset

b. Hand Movement Direction

This dataset was used in the BCI competition where participants perform wrist movement in four different directions and their brain (MEG) activity is recorded. A participant is to move a joystick from the center position to one of the four target classes; left, right, front, or back, with the right hand and wrist. The data from 0.4 s before to

0.6 s after the movement was recorded and resampled at 400Hz. There are 10 series per case from 10 MEG channels.

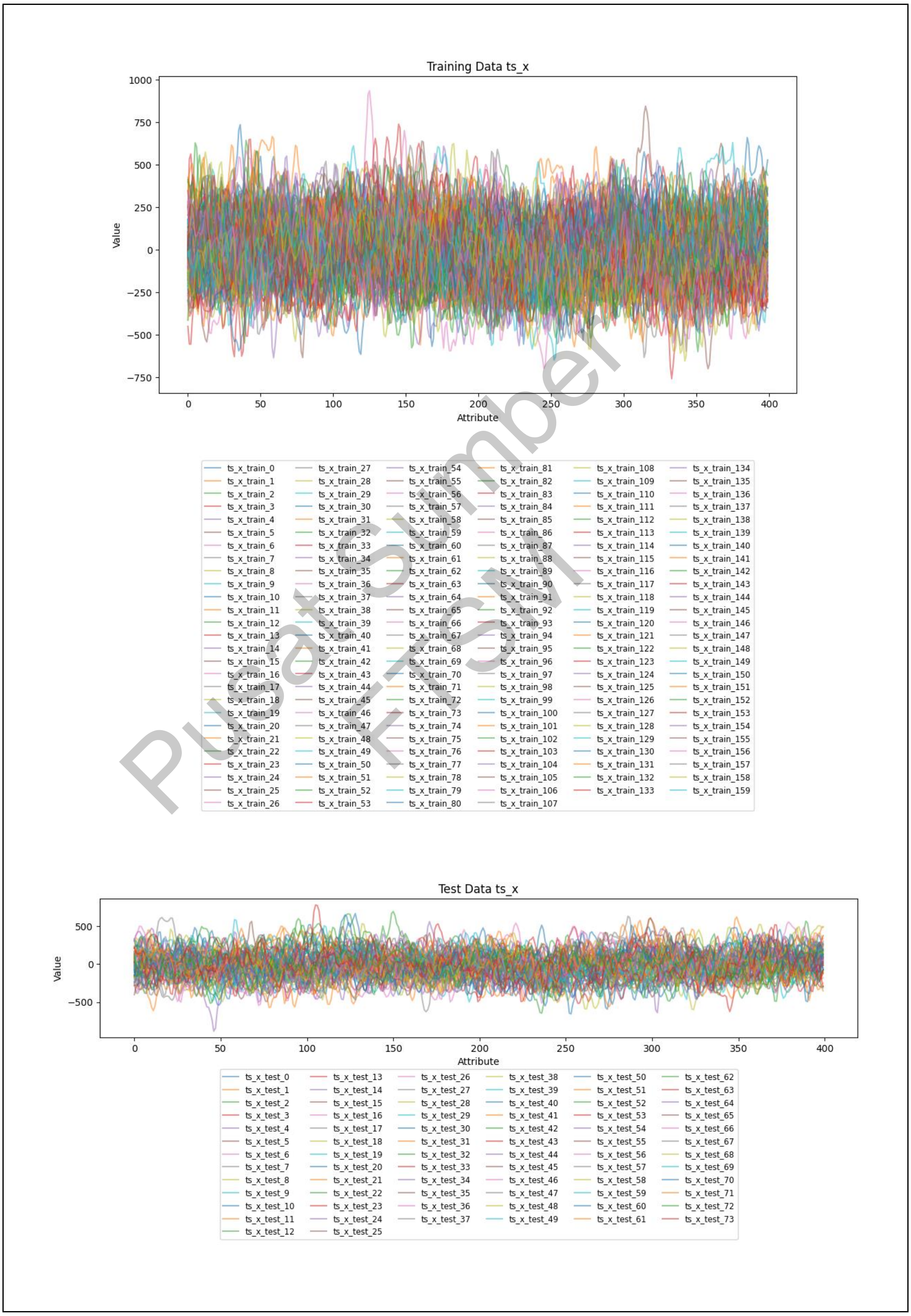
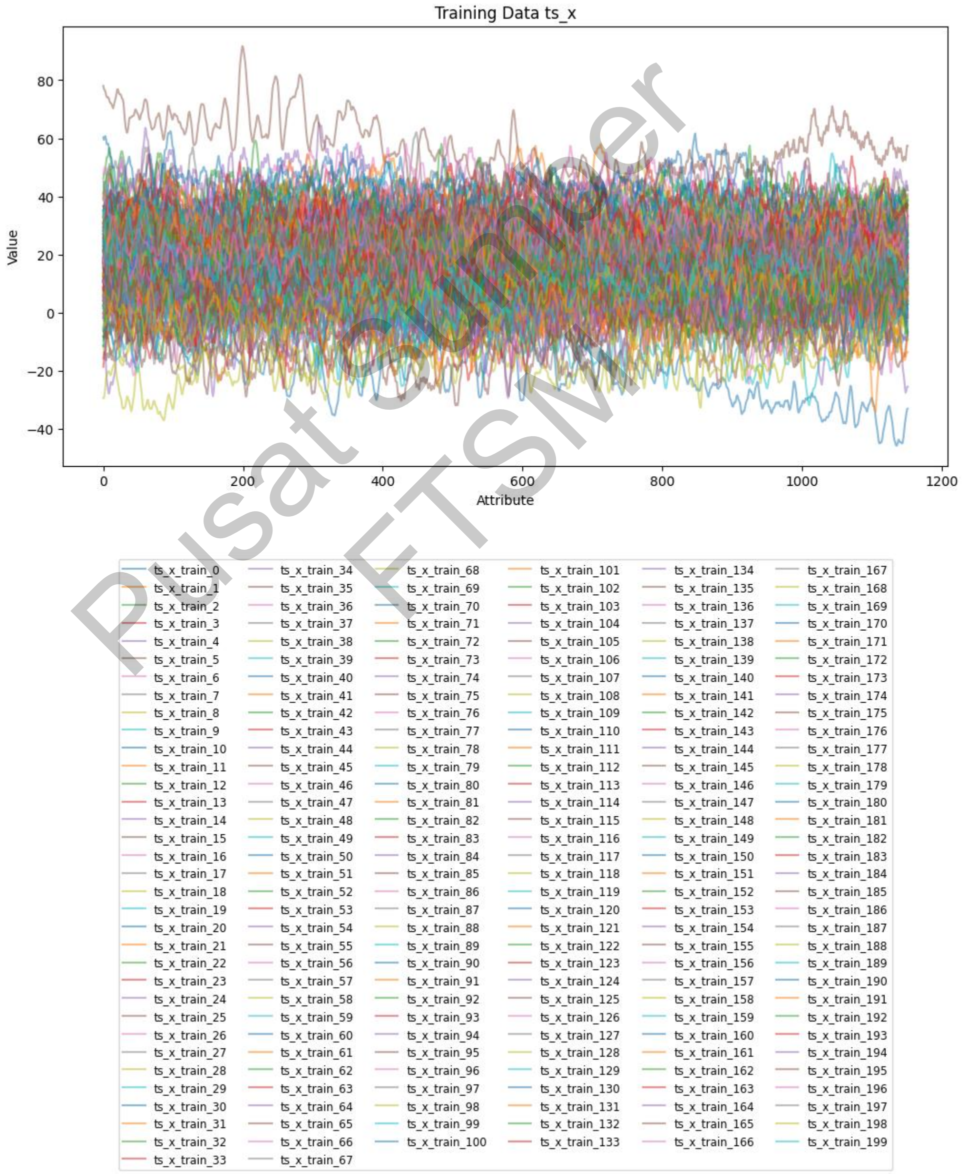


Figure 3.5 Visualization of Hand Movement Direction dataset

c. Self Regulation SCP2

This dataset was used in the BCI II competition. An artificially respirated ALS patient's cortical potentials were recorded when requested to move a cursor up and down on a computer screen. The target class is measured as positive/downward movement on screen and negative/upward movement on screen. The sampling rate of 256Hz and record length of 4.5 s produces 1152 samples in each trial.



to be continued…

…continuation

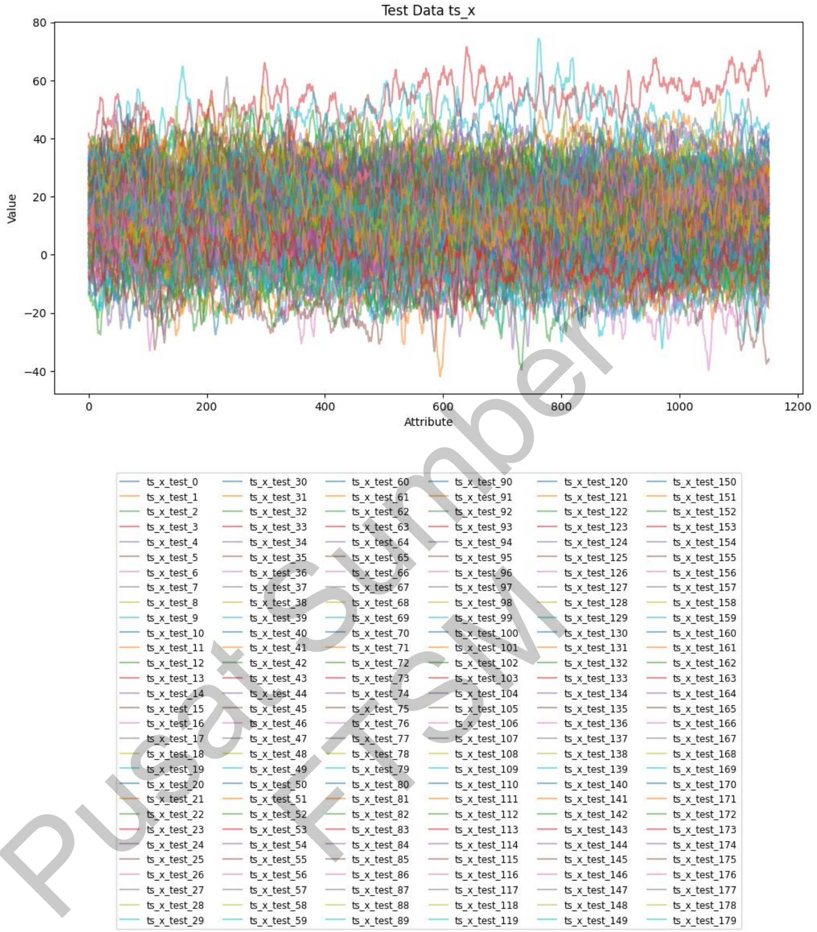


Figure 3.6 Visualization of Self Regulation SCP2 dataset

d. **Stand Walk Jump**(Goldberger et al. 2000)

Four pairs of electrodes built on a patch are placed on a healthy 25-year-old male to monitor the effect of motion artifacts on ECG signals. The classes from this are standing, walking, and jumping.

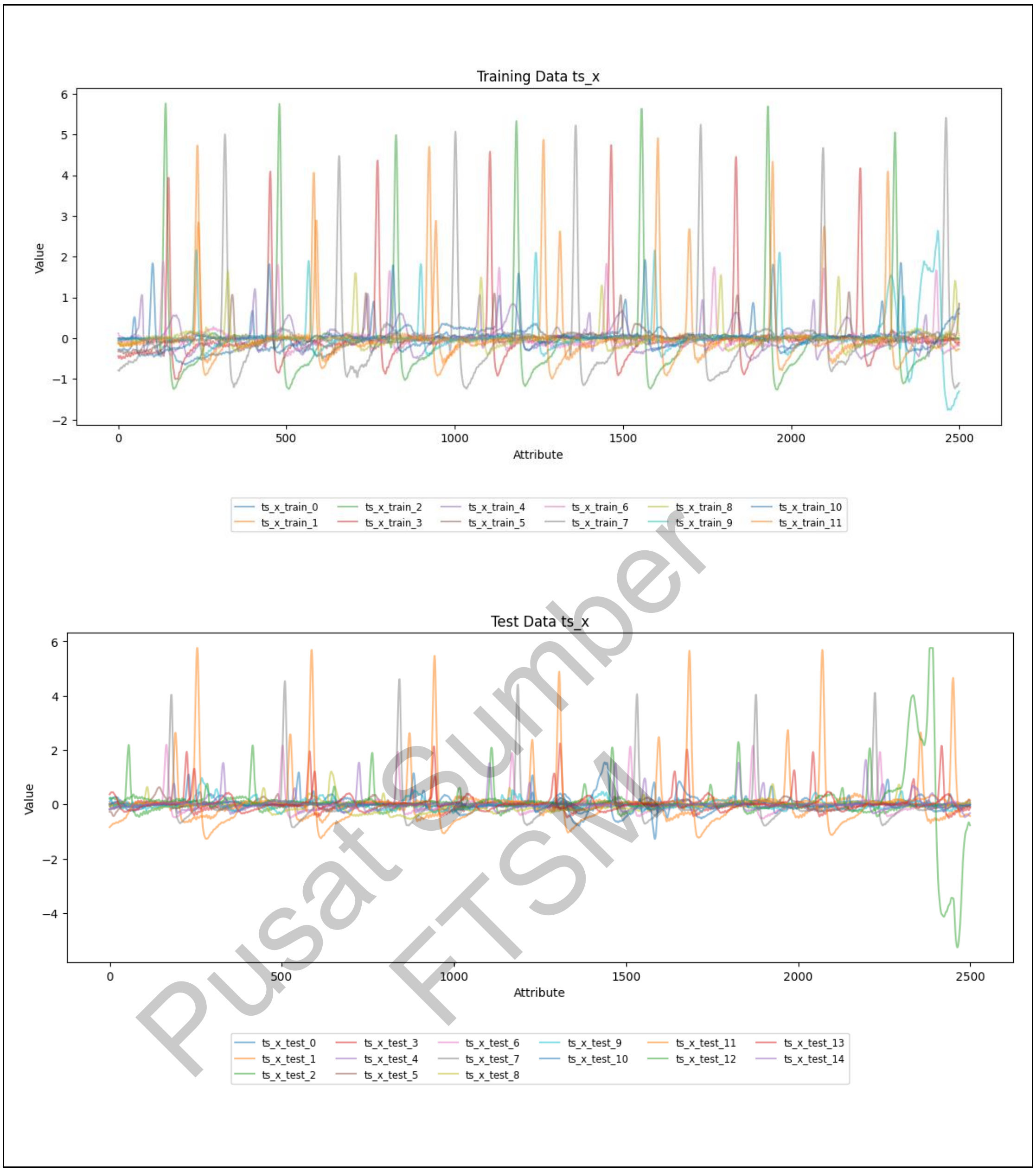


Figure 3.7 Visualization of StandWalkJump dataset

3.3.2 Accelerometer/gyroscope

Accelerometer measures impacts or vibrations by the speed and gyroscopes measure rotational motions. Basic Motions and NATOPS are some of the datasets that uses accelerometers and gyroscopes. These data types are also known as Human Activity Recognition (HAR)

a. Basic Motion

Four students wore a smartwatch each and performed walking, resting, running, and badminton activities which were the classes. They recorded the motion five times and data was sampled once every tenth of a second for ten seconds. Accelerometer x,y,z and gyroscope x,y,z data were collected.

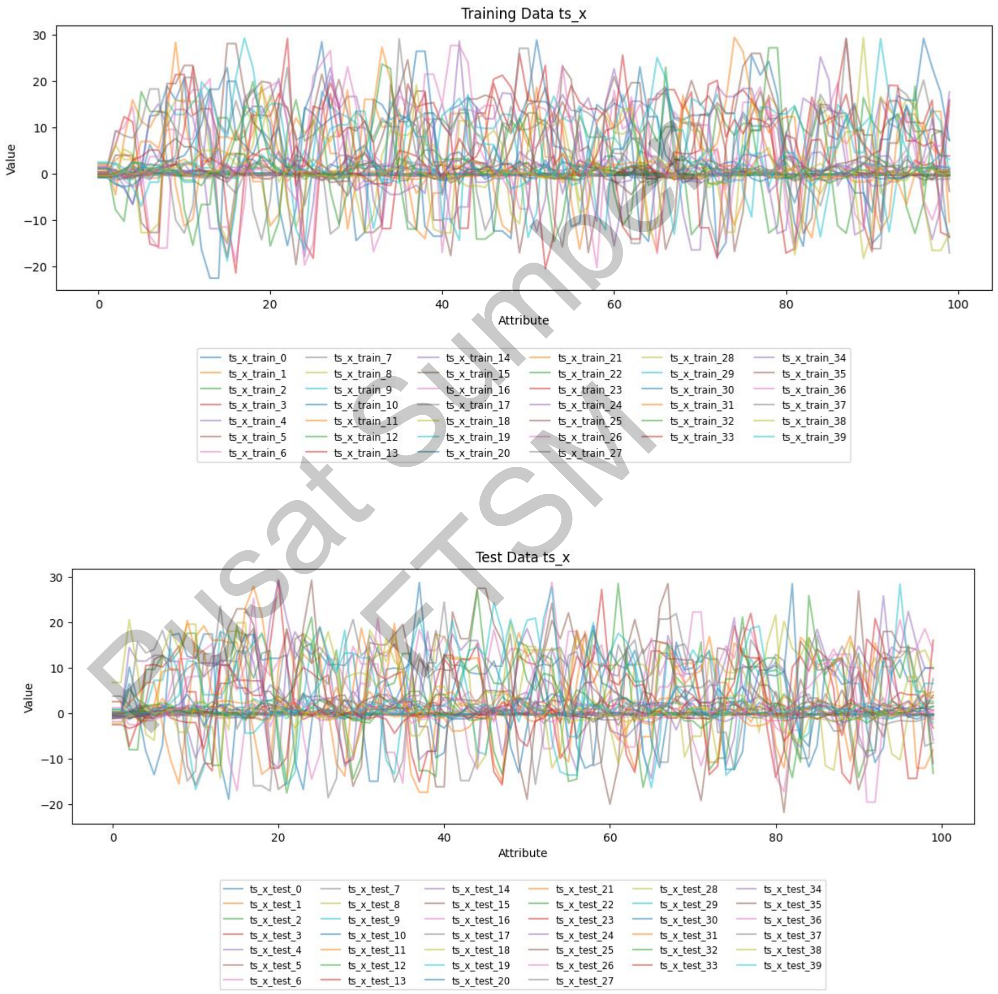
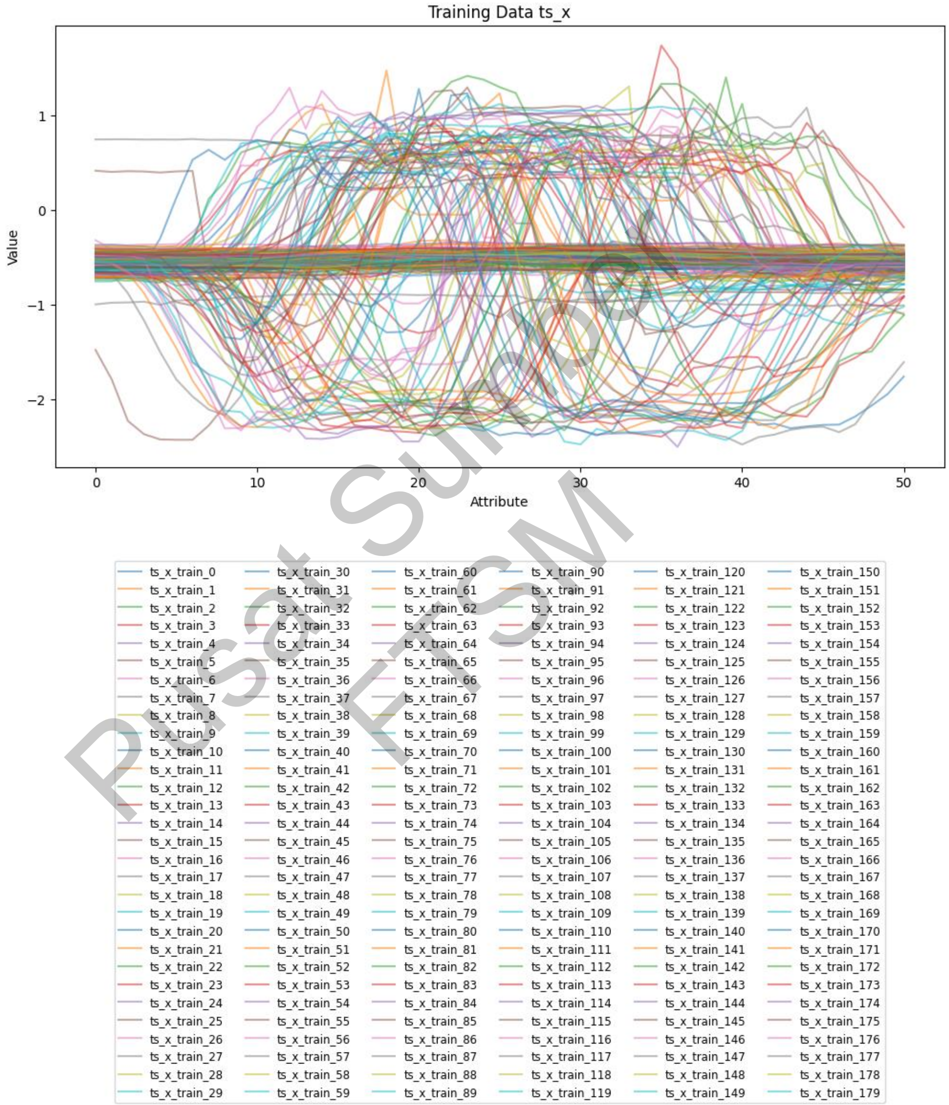


Figure 3.8 Visualization of Basic Motions dataset.

b. **NATOPS** (Ghouaiel et al. 2017)

Data is collected from eight locations on the hands, elbows, wrists, and thumbs and classified into six actions; have command, all clear, not clear, spread wings, fold wings, and lock wings.



to be continued….

…continuation

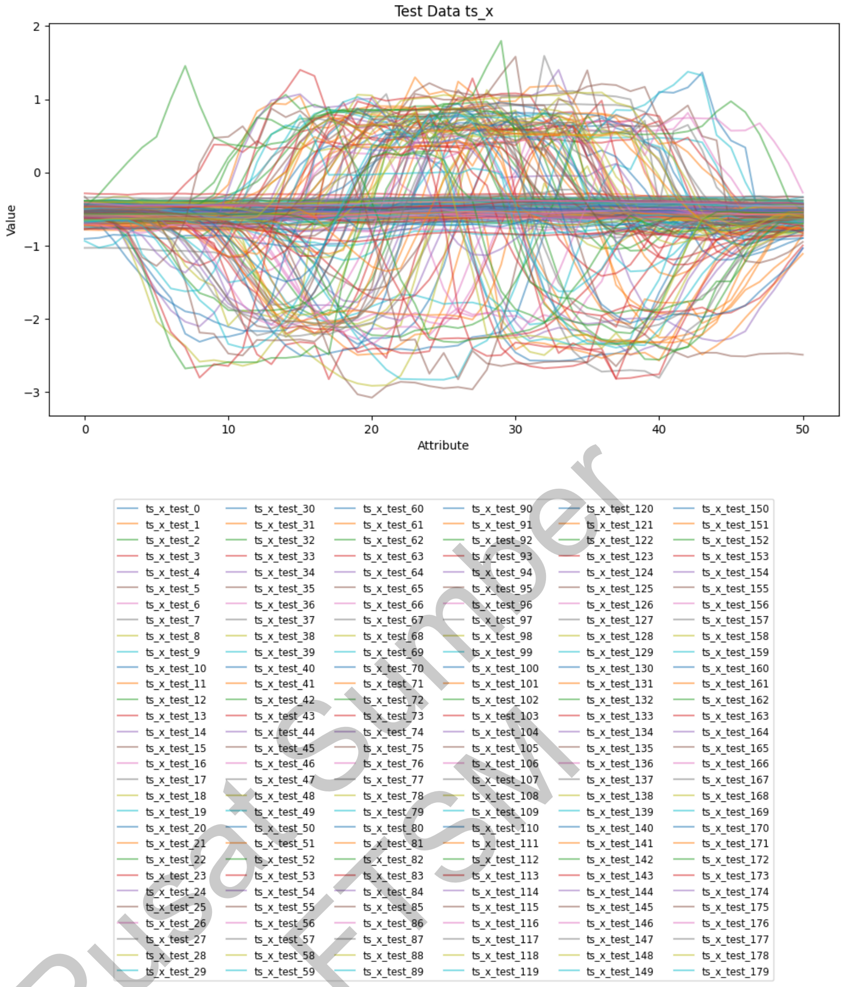


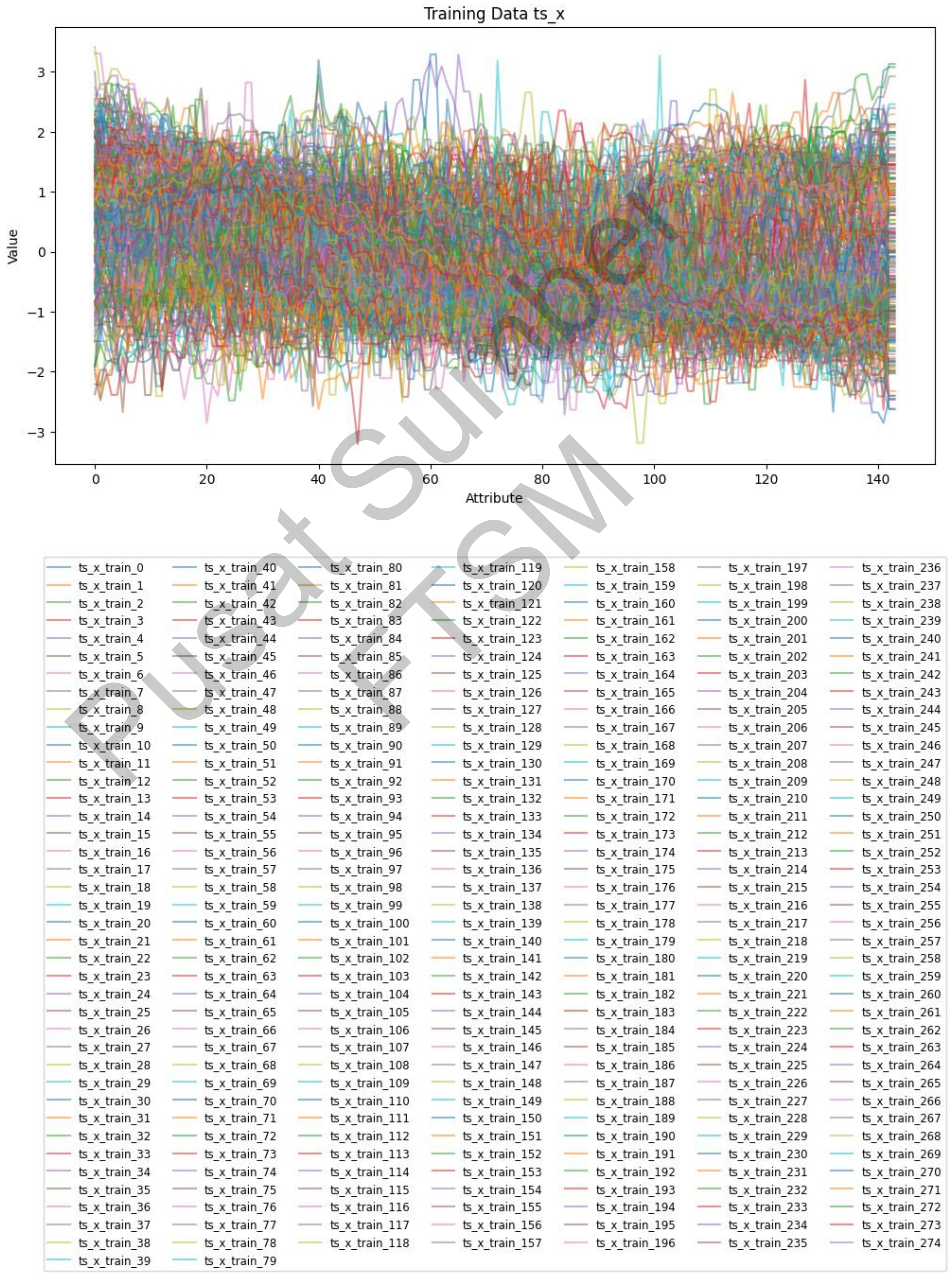
Figure 3.9 Visualization of NATOPS dataset

3.3.3 Coordinates

Coordinates are motion data obtained in Cartesian space such as gesture and digit recognition.

a. Articulary Word Recognition (Wang Jun et al. 2013)

Tongue and lip movement are monitored using an Electromagnetic Articulograph during speech. A number of 25 words from native English speakers are collected as data by attaching head, tongue, lips, and jaw. Nine out of 36 dimensions are used in this data.



to be continued…

…continuation

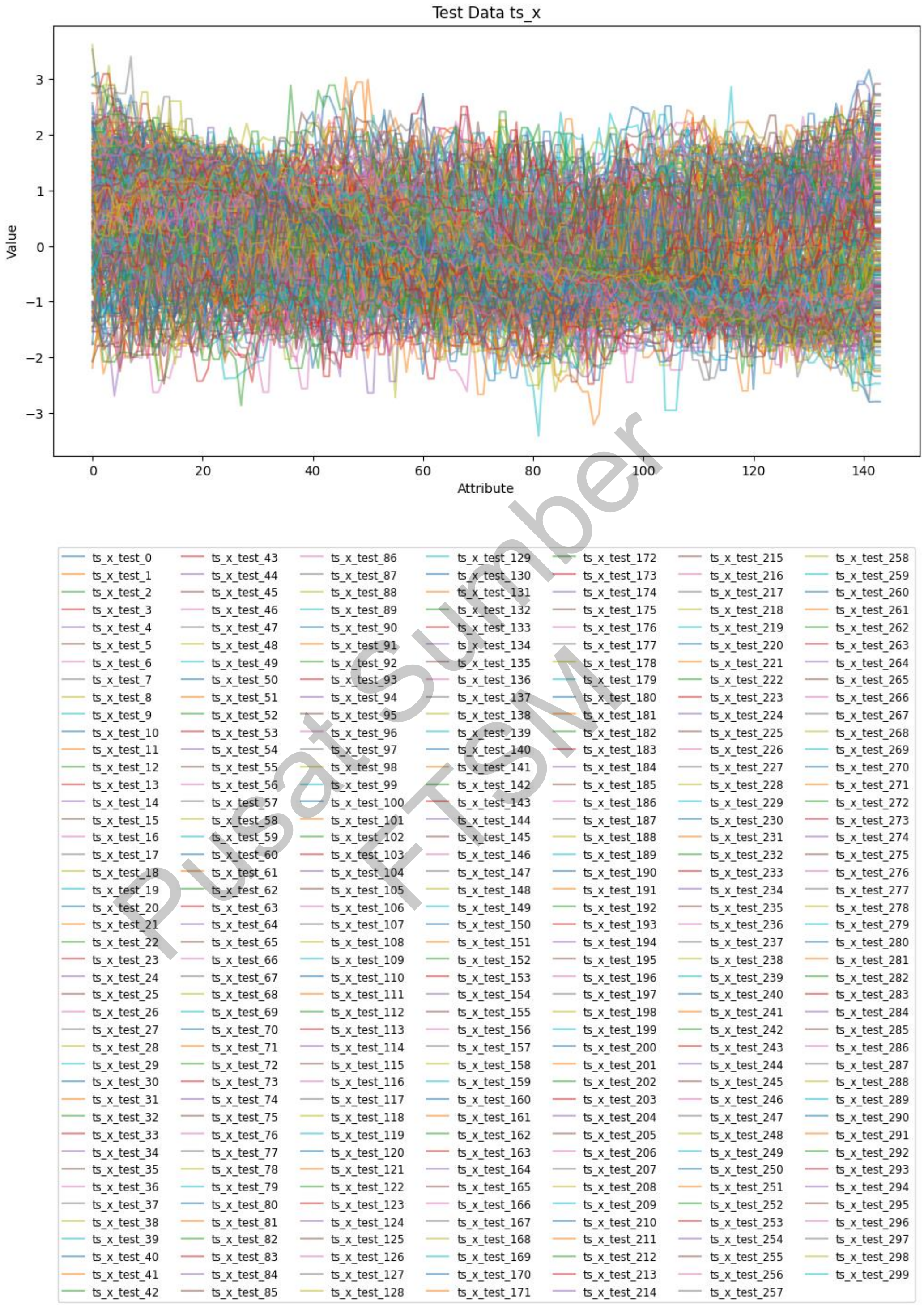


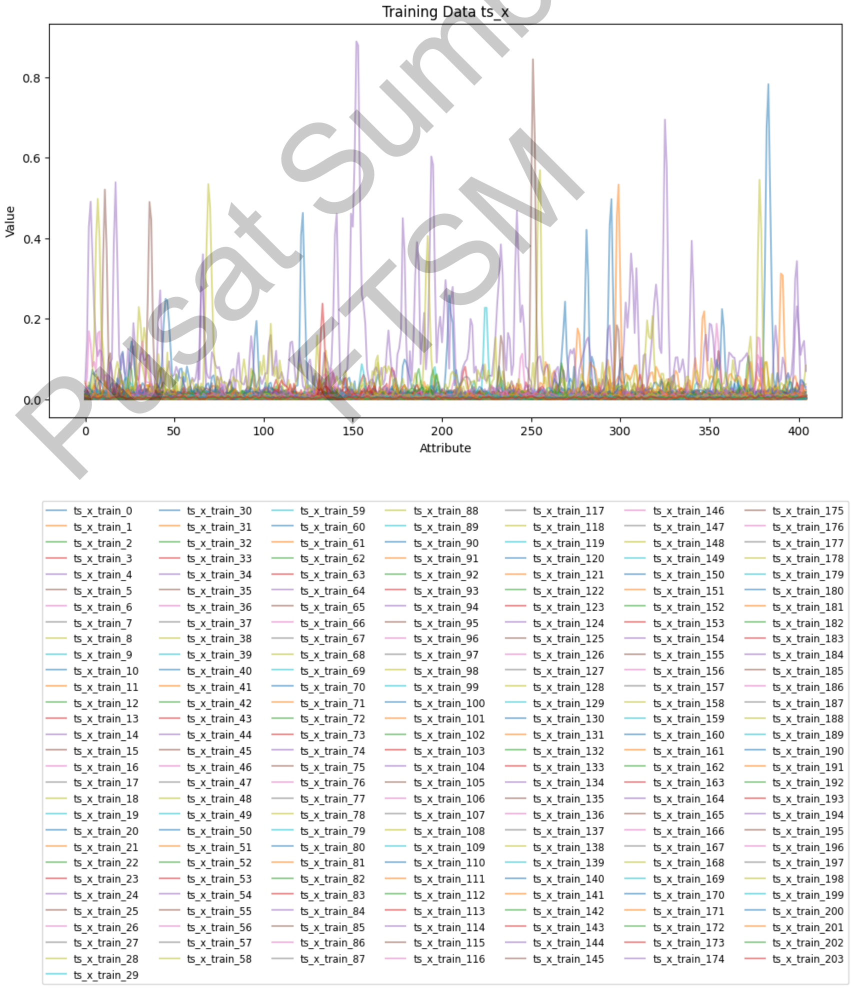
Figure 3.10 Visualization of Articulary Word Recognition dataset

3.3.4 Audio

Audio, or audio spectra, is the best to describe time series data and majority of real-life applications are of this type.

a. **Heartbeat**(Goldberger et al. 2000)

This dataset obtained from PhysioNet/CinC Challenge 2016 records the hearts sound from healthy and pathological patients collected from contributors around the world. Heart sounds are collected from four different locations of the body and classified as normal or abnormal.



to be continued…

…continuation

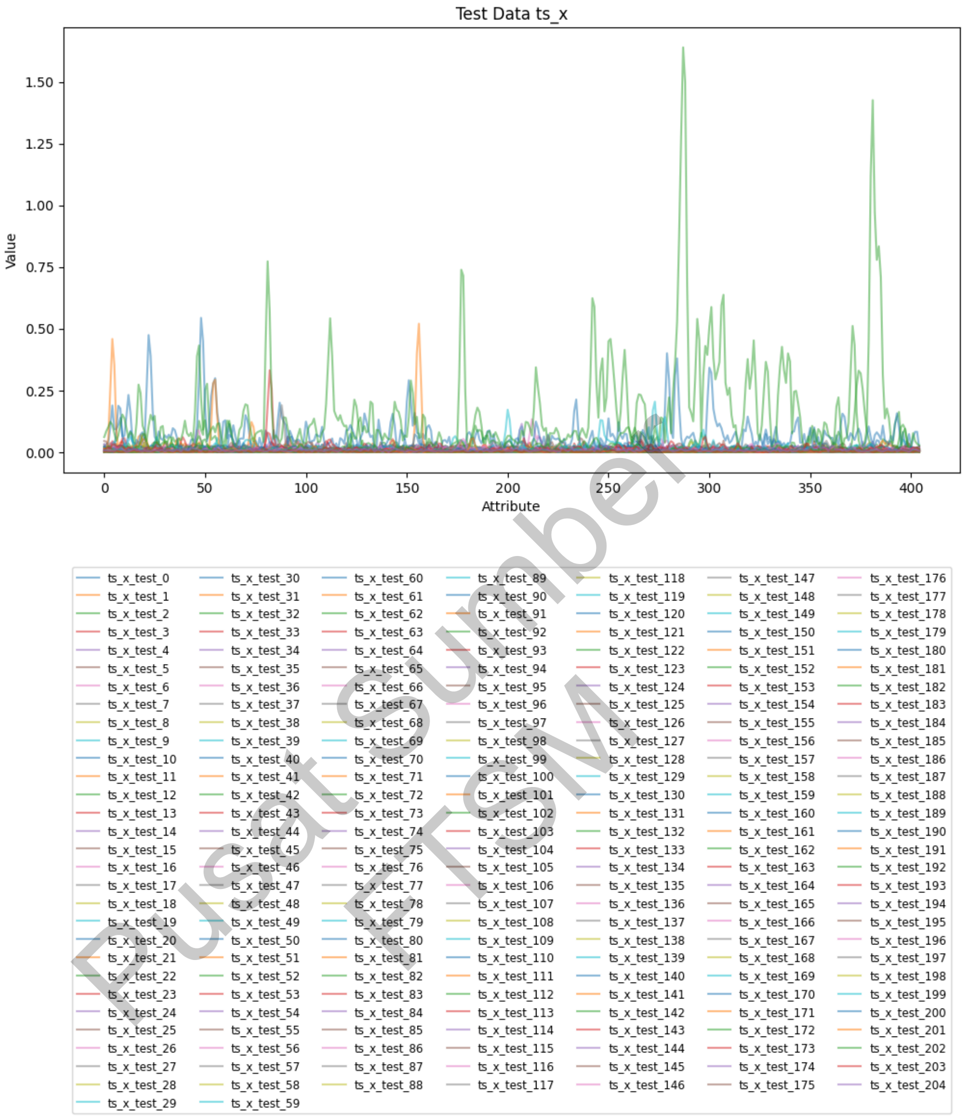


Figure 3.11 Visualization of Heartbeat dataset.

3.4 METHODS

3.4.1 Dynamic Time Warping

a. Euclidean Distance

Initial distance measures use Euclidean distance; which measures the sum of squared distance between two points; for example, a, and b in equation 3.1 (Sammour et al. 2019).

$$d_E(a,b) = \sqrt{\sum_{i=1}^{m}(a_i - b_i)^2} \tag{3.1}$$

Euclidean distance does not consider the order and only allows one to one point comparison whereas DTW allows many to one comparison as shown in Figure 3.12. Therefore, DTW has more advantages due to the elastic distance measure which calculates all the pointwise distance matrix for all the points in the series and selects the minimal cost matrix.

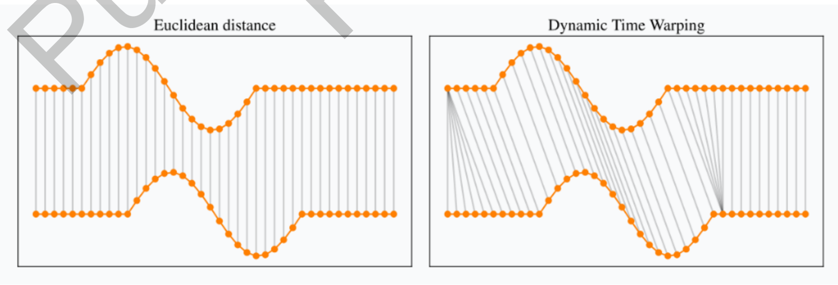


Figure 3.12 Comparison of DTW and ED

Source: Tavenard Romain (2021)

b. Distance in Dynamic Time Warping

The distance for DTW between two series of equal length series A = $(a_1, a_{2,} a_{3,\ldots\ldots\ldots} a_m)$ and series B = $(b_1, b_{2,} b_{3,\ldots\ldots\ldots} b_m)$ can be calculated by equations 3.2 to 3.6 in a stepwise manner (Ruiz et al. 2021).

Step 1
M is a matrix of m x m,where

$$M_{i,j} = \left(a_i - \ b_j\right)^2 \tag{3.2}$$

Step 2
Where warping path

$$P = \left((e_1, f_1,), (e_2, f_2,), \ldots \ldots . (e_s, f_s,),\right) \tag{3.3}$$

is a set of matrix index from M with constraints

$$(e_1, f_1,) = (1,1)$$

$$(e_s, f_s,) = (m, m)$$

$$0 \leq e_{i+1} - \ e_i \ \leq 1 \, for \ all \ i \ < m$$

$$0 \leq f_{i+1} - \ f_i \ \leq 1 \, for \ all \ i \ < m$$

Step 3

Let $p_i = M_{ei,fi}$ be the distance path for $D_{p=} \ \sum_{i=1}^{m} p_i$ (3.4)

Step 4
Find the warping path with the minimum accumulative distance

$$P = min_{p \in P} D_p(A, B) \tag{3.5}$$

Step 5

Calculate optimum distance

$$DTW(i,j) = M_{i,j} + min \begin{cases} DTW \ (i-1, j) \\ DTW \ (i, j-1) \\ DTW \ (i-1, j-1) \end{cases} \tag{3.6}$$

Final distance = *DTW (m,m)*

c. Univariate Time Series Dynamic time warping

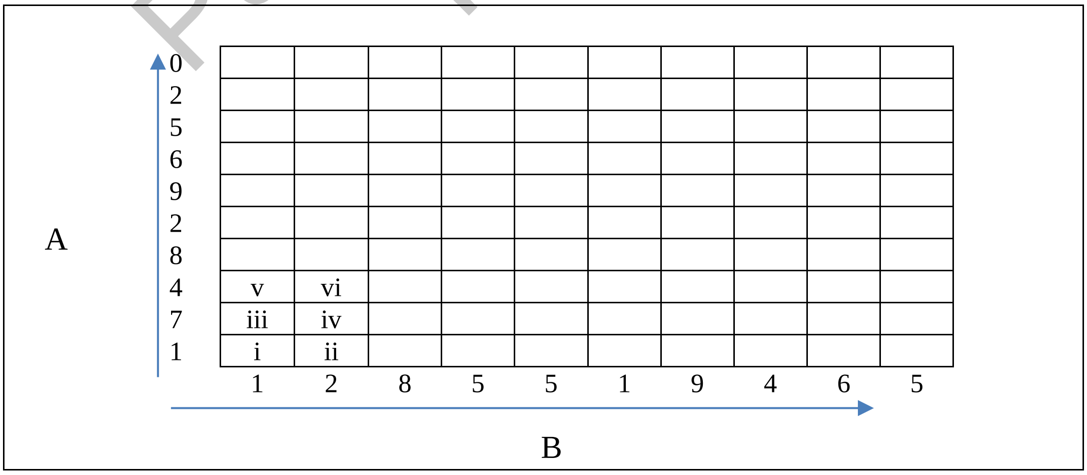
DTW searches for the optimum alignment between time series by calculating the distance between the series. DTW can be performed in univariate or multivariate datasets.(Zhao & Itti, 2017).

An example dataset in Table 3.3 as the data used in this dissertation is too large.

Table 3.3 Example dataset for DTW calculation

Series A	Series B
1	1
7	2
4	8
8	5
2	5
9	1
6	9
5	4
2	6
0	5

Step 1
Create a cost matrix where series A on the left is labelle**d** from bottom to top, and series B on the bottom is labelle**d** from left to right (Salvador & Chan 2007).



A										
0										
2										
5										
6										
9										
2										
8										
4	v	vi								
7	iii	iv								
1	i	ii								
	1	2	8	5	5	1	9	4	6	5
	B									

Figure 3.13 Create a cost matrix

Step 2
Calculate the cost matrix with equation 3.6 and fill in calculated figures starting from the bottom left corner.

Calculation according to equation 3.6 for

Value i	: |1 –1| + min(0)	= 0
Value ii	: |1 –2| + min(0)	= 1
Value iii	: |7 -1| + min(0)	= 6
Value iv	: |7 -2| + min(6, 0, 1)	= 5
Value v	: |4 -1| + min(6)	= 9
Value vi	: |4 -2| + min(9,6,5)	= 7

Calculation for all the distances:

A

0	36	29	32	22	22	16	24	16	18	17
2	35	27	24	17	17	15	21	12	14	12
5	34	27	18	14	14	18	14	10	10	9
6	30	24	15	14	14	18	10	9	9	10
9	25	20	13	13	13	15	7	12	13	17
2	17	13	12	9	9	7	14	10	14	17
8	16	13	6	6	6	11	8	12	14	17
4	9	7	6	3	4	7	12	12	14	15
7	6	5	2	4	6	12	14	17	18	20
1	0	1	8	12	16	16	24	27	32	36
	1	2	8	5	5	1	9	4	6	5

B

Figure 3.14 Calculation for all the distances.

Step 3
Determination of the warping path, starting from the top right to the bottom left corner, using the smallest figure to find the least warp distance for the time series A and B (Salvador & Chan 2007).

A

0	36	29	32	22	22	16	24	16	18	17
2	35	27	24	17	17	15	21	12	14	12
5	34	27	18	14	14	18	14	10	10	9
6	30	24	15	14	14	18	10	9	9	10
9	25	20	13	13	13	15	7	12	13	17
2	17	13	12	9	9	7	14	10	14	17
8	16	13	6	6	6	11	8	12	14	17
4	9	7	6	3	4	7	12	12	14	15
7	6	5	2	4	6	12	14	17	18	20
1	0	1	8	12	16	16	24	27	32	36
	1	2	8	5	5	1	9	4	6	5

B

Warping path: [17,12,9,9,9,7,7,6,3,2,1,0]

Figure 3.15 Determination of warping path.

Step 4
Final Distance Calculation with equation 3.7 (Herrmann et al. 2023)

$$D = {\sum_{i=1}^{l} d\,(i)} \Big/ {\sum_{i=1}^{l} l} \tag{3.7}$$

where l is the length of the series d.

From above, l = 12

DTW distance = (17+12+9+9+9+7+7+6+3+2+1+0)/12

= 6.833

The example shows the DTW for the univariate dataset.

d. Multivariate distances

DTW for univariate applies to multivariate data with differences in distance calculation. There are two different calculations which are independent and dependent DTW.

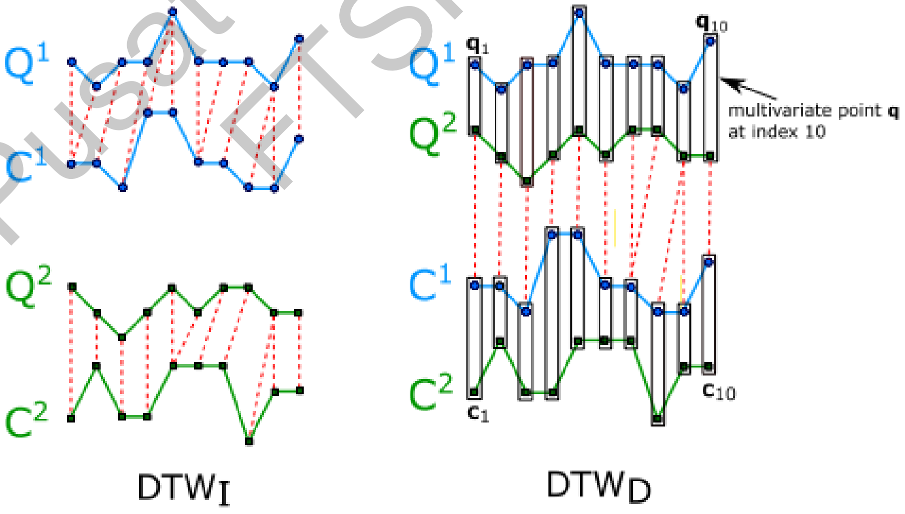


Figure 3.16 Independent DTW on left and dependent DTW on right. Dimension A is in blue and dimension B is in green.

Source: Shifaz et al. (2023)

For independent DTW, the distance between each channel is calculated independently (similar to univariate), and then total of each distance is added together as shown in equation 3.8 (Shokoohi-Yekta et al. 2017). In equation 3.8, each channel is independent to others.

$$d_{independent}(a, b) = \sum_{c=1}^{k} d(a_c, b_c) \tag{3.8}$$

where k is the number of channels and

d (a_c, b_c) is defined as distance of cth channel of a and cth channel of b.

Example calculation for independent DTW of two points a and b with three channels using Equation 3.8:

Point a = (1, 2, 3),

Point b = (4, 5, 6)

Distance 1: $d_{independent} = |1 - 4| = 3$
Distance 2: $d_{independent} = |2 - 5| = 3$
Distance 3: $d_{independent} = |3 - 6| = 3$

DTW distance $_{\text{independent}}$ = 3+3+3 = 9

For dependent DTW, the distance of all the matrix, M data points are calculated as cumulative squared Euclidean distance instead of single points as shown in equation 3.9 (Shokoohi-Yekta et al. 2017). In this dissertation, we use DTW from aeon toolkit[2] which assumes the dependent approach.

$$M_{i,j}(a, b) = \sqrt{\sum_{c=1}^{k} d(a_{c,i} - b_{c,j})^{2}} \tag{3.9}$$

where $a_{c,i}$ is the ith data point in cth channel

and $b_{c,j}$ is the jth data point in cth channel.

[2] https://www.aeon-toolkit.org/en/v0.9.0/

Point a = (1, 2, 3),

Point b = (4, 5, 6)

Distance : $d_m = \sqrt{(1-4)^2 + (2-5)^2 + (3-6)^2}$

$= \sqrt{9+9+9}$

$= 5.2$

DTW distance $_{\text{dependent}}$ = 5.2

The calculation example showed that independent DTW and dependent DTW produce different outputs. DTW used in this dissertation is the dependent DTW approach and an example of implementation will be based on this.

e. Multivariate Time Series Dynamic Time Warping

Section 3.4.1(b) shows how the DTW distance of a two-time series is calculated for univariate data. For multivariate data, the principle of extending the univariate method to multivariate is performed in step 2 and step 4 below.

In step 2, by incorporating the dependent DTW formula from Equation 3.9 in Section 3.4.1(d) into Equation 3.6 in Section 3.4.1(b), the matrix is updated with new values for determination of the warping path with the lowest distance value.

In step 4, distance calculation for univariate data in Equation 3.7 in section 3.4.1(c) is also modified by using Equation 3.9. Equation 3.9 which is the calculation of distance for dependent DTW is used to replace the univariate DTW distance calculation.

Example dataset for multivariate time series:

Sequence A = [(1,2,1), (2,3,2), (3,4,3)]

Sequence B = [(1,1,2), (2,2,3), (3,3,4)]

Step 1

Create cost matrix with an extra row and column for initialisation.

	∞	vii	viii	ix
A	∞	iv	v	vi
	∞	i	ii	iii
		∞	∞	∞
			B	

Figure 3.17 Create cost matrix.for multivariate time series.

Step 2

Calculate the cost matrix by incorporating equation 3.9 into equation 3.6 and fill in calculated figures starting from the bottom left corner.

To calculate value i (*i* = 1, *j* = 1):

Cost $= \sqrt{(1-1)^2 + (2-1)^2 + (1-2)^2} = \sqrt{2} = 1.41$

Value i = 1.41 + min (∞, 0, ∞) = 1.41

To calculate value ii (*i* = 1, *j* = 2):

Cost $= \sqrt{(1-2)^2 + (2-2)^2 + (1-3)^2} = \sqrt{5} = 2.24$

Value ii = 2.24 + min (1.41, ∞, ∞) = 3.65

To calculate value iii (*i* = 1, *j* = 3):

Cost $= \sqrt{(1-3)^2 + (2-3)^2 + (1-4)^2} = \sqrt{14} = 3.74$

Value iii = 3.74 + min (3.65, ∞, ∞) = 7.39

To calculate value iv (*i* = 2, *j* = 1):

Cost $= \sqrt{(2-1)^2 + (3-1)^2 + (2-2)^2} = \sqrt{5} = 2.24$

Value iv = 2.24 + min (∞, ∞, 1.41) = 3.65

To calculate value v (*i* = 2, *j* = 2):

Cost $= \sqrt{(2-2)^2 + (3-2)^2 + (2-3)^2} = \sqrt{2} = 1.41$

Value v = 1.41 + min (3.65, 1.41, 3.65) = 2.82

To calculate value vi (*i* = 2, *j* = 3):

Cost $= \sqrt{(2-3)^2 + (3-3)^2 + (2-4)^2} = \sqrt{5} = 2.24$

Value vi = 2.24 + min (2.82, 3.65, 7.39) = 5.06

To calculate value vii (*i* = 3, *j* = 1):

Cost $= \sqrt{(3-1)^2 + (4-1)^2 + (3-2)^2} = \sqrt{14} = 3.74$

Value vii = 3.74 + min (∞, ∞, 3.65) = 7.39

To calculate value viii (*i* = 3, *j* = 2):

Cost $= \sqrt{(3-2)^2 + (4-2)^2 + (3-3)^2} = \sqrt{5} = 2.24$

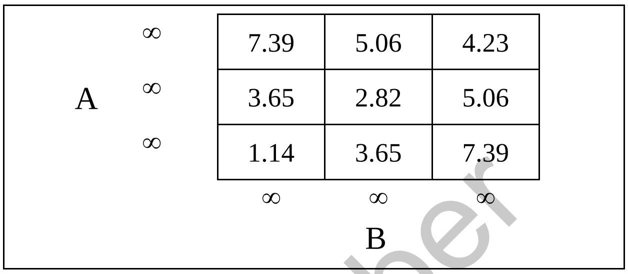
Value viii = 2.24 + min (7.39, 3.65, 2.82) = 5.06

To calculate value ix (*i* = 3, *j* = 3):

Cost $= \sqrt{(3-3)^2 + (4-3)^2 + (3-4)^2} = \sqrt{2} = 1.41$

Value ix = 1.41 + min (5.06, 2.82, 5.06) = 4.23

Calculation for all the distances:

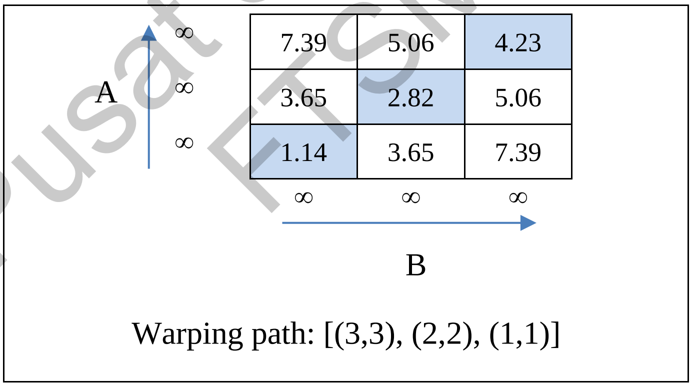


	∞	7.39	5.06	4.23
A	∞	3.65	2.82	5.06
	∞	1.14	3.65	7.39
		∞	∞	∞
			B	

Figure 3.18 Calculation of all the distances for multivariate time series data.

Step 3

Determination of warping path starting from the top right to bottom left corner via the smallest value:



	∞	7.39	5.06	4.23
A	∞	3.65	2.82	5.06
	∞	1.14	3.65	7.39
		∞	∞	∞
			B	

Warping path: [(3,3), (2,2), (1,1)]

Figure 3.19 Determination of warping path for multivariate time series data.

Step 4

Final Distance Calculation by incorporating dependent DTW of equation 3.9 into equation 3.7:

DTW

$$= \text{Distance of } [(3,3),(2,2),(1,1)] / 3$$

$$= \frac{\sqrt{(3-3)^2+(4-3)^2+(3-4)^2} + \sqrt{(2-2)^2+(3-2)^2+(2-3)^2} + \sqrt{(1-1)^2+(2-1)^2+(1-2)^2}}{3}$$

$$= \frac{\sqrt{2} + \sqrt{2} + \sqrt{2^{\square}}}{3}$$

$= 1.41$

f. Shape Dynamic Time Warping

ShapeDTW first uses shape descriptors to represent the temporal points, then aligns different sequences according to the descriptors by DTW. This allows more focus on local shape during the matching process which produces more semantically meaningful results (Zhao & Itti 2017) compared to DTW which finds global optimum. Piecewise aggregate approximation, PAA is obtained by splitting the series into a designated number of intervals,

There are two steps in shaping DTW implementation. Firstly, the original time series is transformed into shape descriptors that represent the local sequence information. The series is divided into parts of sequences. DTW is used to derive the shape of each divided sequence to produce the characteristic descriptor.

Secondly, DTW is used to align these descriptors which is similar to the typical DTW to produce the warping path. Figure 3.22 shows an example of ShapeDTW alignment on the SRS2 dataset.

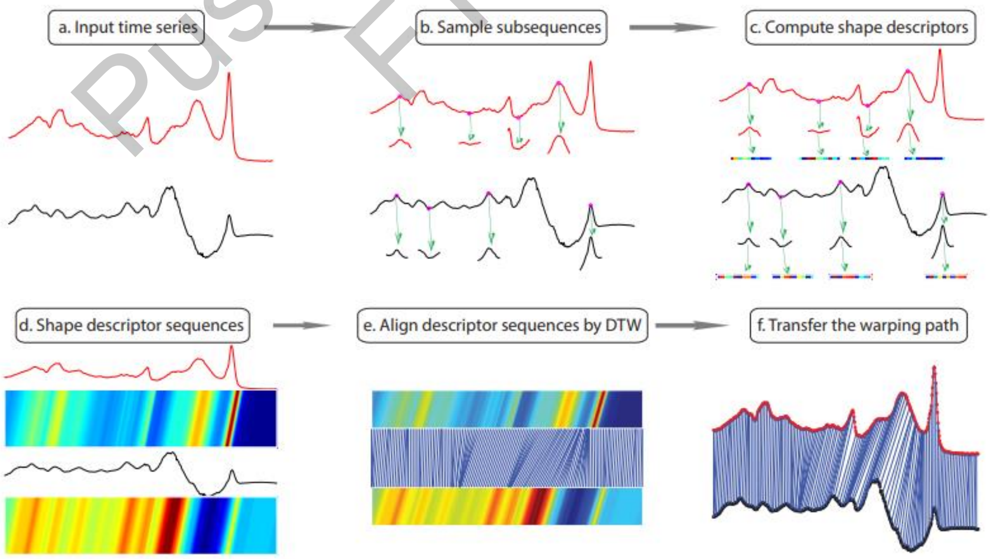

a. Input time series
b. Sample subsequences
c. Compute shape descriptors
d. Shape descriptor sequences
e. Align descriptor sequences by DTW
f. Transfer the warping path

Figure 3.20 ShapeDTW alignment procedure

Source: Zhao & Itti (2017)

```
Shape Dynamic Time Warping
Inputs:
Time series P ∈ $R^{LP}$ and Q ∈ $R^{LQ}$ ; subsequence length $l$; shape descriptor function $F$
ShapeDTW:
    1. Sample subsequences: $S^P$ ← P, $S^Q$ ← Q;
    2. Encode subsequences by shape descriptors:
       $d^P$ ← $F(S^P)$, $d^Q$ ← $F(S^Q)$;
    3. Align descriptor sequences $d^P$ and $d^Q$ by DTW.
Outputs:
  Warping matrices: $W^P$ and $W^Q$;
  ShapeDTW distance: k W˜ P ∗ · d P − W˜ Q ∗ · d Q k
```

Figure 3.21 Pseudo code of Shape Dynamic Time Warping

Source: Zhao & Itti (2017)

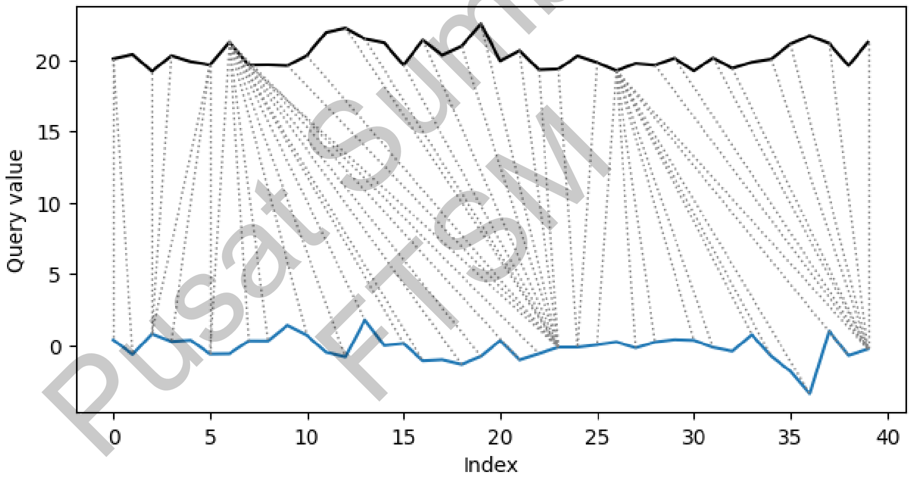


Figure 3.22 Example of Shape DTW alignment of SRS2 dataset

ShapeDTW was initially performed on univariate time series data. In this dissertation, the multivariate data is converted to one-dimensional for further analysis

3.4.2 Dual Attention Network

The proposed baseline framework of DA-Net is shown in Figure 3.23. It is a hierarchical structure with four time-block partition layers and dual-attention blocks.

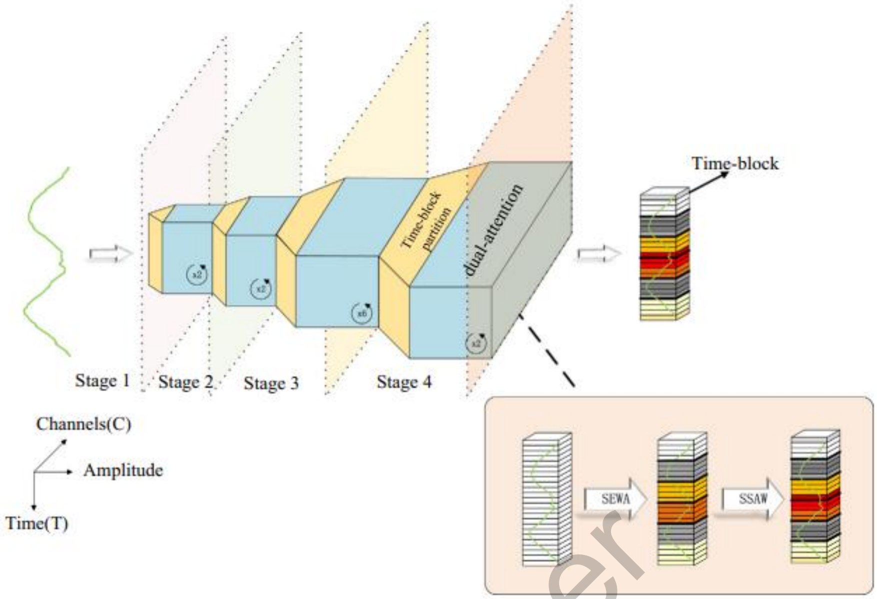


Figure 3.23 The overall framework of DA-Net

Source: Chen et al. (2022)

The time-block partition layers reduce time series length by concatenating the four non-overlap neighbor timestamps as a time-block (a token in NLP), and each time-block flattened and projected to 4C-dimensional embedding. This is performed at the beginning of each stage and reduces length of input data by a factor of four (Figure 3.24).

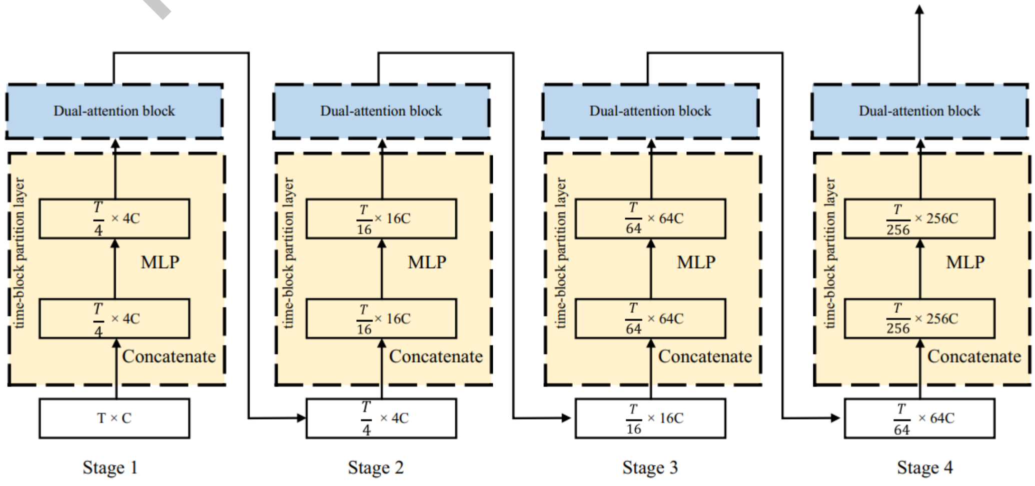
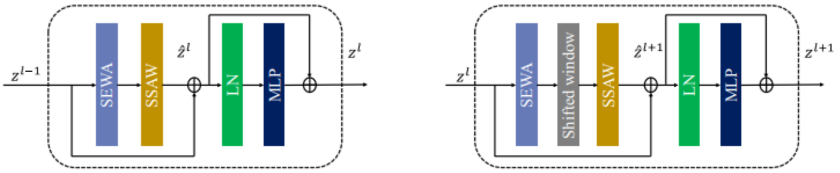


Figure 3.24 Attribute transformation of DA-Net

Source: Chen et al. (2022)

Attention blocks are made of SEWA layer, SSAW layer, Layer Normalisation (LN) layer, MLP layer as shows in Figure 3.25. The second module has an additional ‘shifted window layer’ which shifts time-blocks within the window. The benefit of this is to resolve long-time dependencies restricted to local window partitioning.



(a) The first module of dual-attention block

(b) The second module of dual-attention block

Figure 3.25 Dual-attention block

Source: Chen et al. (2022)

The SEWA layer focuses on local distinguishing attributes with the aid of global window-attribute. It divides attributes by windows $\mathrm{X} \in \mathrm{R}^{M\ \mathrm{x}\ C\ \mathrm{x}\ .W}$ where M = the number of windows, C = the number of channels, and W = time block within a window. By squeeze and excitation, it suppresses non-significant windows and amplify significant ones and normalise weights to zero to one using sigmoid function. This step enables high-level attributes of MTS to embed into a window as it calculates the values from time blocks and channels.

Sparse Self-Attention within the Window (SSAW) layer captures the global long-range dependencies, reduces computation complexity by Sparse-attention, and expands window size.

Hyperparameters in DA-Net baseline experiment are set with batch size B = 16, window size M = 64, channel number of hidden layer C = 96, multi-head numbers = {3,6,12,6} and layer numbers = {2,2,6,2} (Chen et al. 2022).

a. SEWA layer

SEWA layer divide attributes by windows into X $\in$ R$^{M \text{ x } C \text{ x } .W}$ where M = number of windows, C = number of channels and W = number of time block within a window. This input X is then transformed and mapped to attribute S = R$^{M\prime \text{ x } C\prime \text{ x } W\prime}$ after squeeze and excitation in equation 3.10 to 3.12 (Chen et al. 2022).

Step 1: Squeeze

$$Z = F_{sq}\ (X) = \frac{1}{C\ x\ W} \sum_{c}^{C} \sum_{w}^{W} X\ (c, w) \tag{3.10}$$

Step 2: Excitation

$$\mathrm{H} = F_{ex}\ (Z) = W_2 ReLU(W_1 Z) \tag{3.11}$$

where W$_1$ and W$_2$ = learning parameters of linear projections

$$\mathrm{S} = F_{scale}\ (H, X) = \mathrm{Xsigmod(H)} \tag{3.12}$$

b. SSAW layer

Figure 3.27 shows a summary of how SSAW selects the top *u*. In this case, *u* is set as two. The red points are query and key lines with huge difference whereas the green points are no significant difference. The output is the concatenated values of the top *u* scores and mean values.

Sparse Self Attention within Windows

```
Input  Queries, $Q \in R^{M_Q\,x\,C}$ , Keys, K∈ $R^{M_K\,x\,C}$;  and Values, V∈ $R^{M_V\,x\,C}$,
$u = M_v \ln M_Q$, $U = M_Q \ln M_K$ ;
where $M_Q$, $M_k$ and $M_v$ = window size
Ensure: Self-attention feature map S;
1: Randomly select $U$ keys as K;
2: Calculate measurement $M$ using $K$ by formulation:
    $\bar{M}(q_i, K) = \max_j \left\{ \frac{q_i k_j^T}{\sqrt{d}} \right\} - \frac{1}{L_K} \sum_{j=1}^{L_K} \frac{q_i k_j^T}{\sqrt{d}}$
3: Select top-$u$ $Q$ using $M$ by formulation:
    $\bar{Q} = top_u \bar{M}$
4: Calculate self-attention feature map $S$ using Q by formulation:
    $S = \begin{cases} Softmax\left(\frac{\bar{Q}K^T}{\sqrt{d}}\right).V \;\; if\ i - th\ query\ is\ top - u\ queries \\ Mean\ (V)\ if\ i - th\ query\ is\ not\ top - u\ queries \end{cases}$
```

Figure 3.26 Pseudo code for Sparse Self Attention within Windows

Source: Chen et al. (2022)

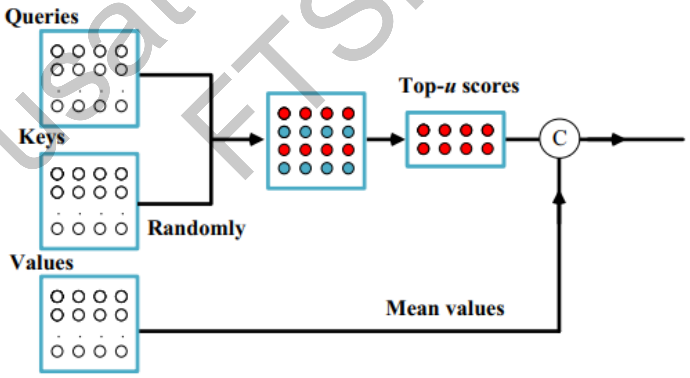


Figure 3.27 Figure of SSAW

Source: Chen et al. (2022)

3.5 EVALUATION

3.5.1 Performance

a. Accuracy, precision, recall, and F1-score

The performance of DTW-DA-Net is evaluated by comparing accuracy, precision, recall, and F1-score. Accuracy is the ability to correctly predict positive and negative values, precision is the ability to correctly predict positive values, recall is the measurement of ability to predict positive cases correctly, and F-measure is the weighted average of precision and recall (Melhem et al. 2021; Sisodia & Sisodia 2018). Formula interpreting the statements are shown in equations 3.13, 3.15, 3.16, and 3.16 (Melhem et al. 2021; Sisodia & Sisodia 2018).

$$\text{Accuracy} = \frac{True\ positive{+}True\ Negative}{(True\ Positive{+}False\ Positive{+}True\ Negative{+}False\ Negative)} \tag{3.13}$$

$$\text{Precision} = \frac{True\ positive}{(True\ Positive{+}False\ Positive)} \tag{3.14}$$

$$\text{Recall} = \frac{True\ positive}{(True\ Positive{+}False\ Negative)} \tag{3.15}$$

$$\text{F1-score} = 2\ \text{x}\ \frac{(Precision\ x\ Recall)}{(Precision{+}Recall)} \tag{3.16}$$

b. Performance evaluation

Comparison of performance of two methods is measured by increase in percentage.

Performance comparison:

$$\frac{\boldsymbol{Accuracy\ of\ Method\ A - Accuracy\ of\ Method\ B}}{\boldsymbol{Accuracy\ of\ Method\ B}}\ \text{x}\ 100 \tag{3.17}$$

3.5.2 Statistical Evaluation

To compare the algorithm's effectiveness on different datasets, the Friedman test is performed with α of 0.05 to test the hypothesis if there is a significant difference in performance. Then, the Wilcoxon signed-rank test is performed with Holm’s alpha correction as performed by Ismail Fawaz et al. (2019). Ruiz et al. (2021)and Chen et al. (2022) also performed the Wilcoxon signed rank test in their MTSC. A critical difference diagram will show the ranking of the best performer to the weakest performer and if there is a statistical difference.

3.6 CHAPTER SUMMARY

This chapter gives an outline of the structure of this research. The research design explains the proposed method to address the research questions and objectives. The datasets are categorized according to different types and will be presented as such in the following chapters. The methods section explains the details of the algorithm structure with related examples. Evaluation methods to be used in the next chapter are explained.

The following chapter gives the results of an implementation of the design explained in this chapter.

CHAPTER IV

RESULTS AND DISCUSSION

4.1 INTRODUCTION

This chapter presents the results and discussion of the proposed DTW-DA-Net and ShapeDTW and evaluation if the proposed methods can meet the research objective. The results and discussion are presented in five sections.

Section 4.2 assesses the first objective "*To assess the applicability of univariate distance-based classification methods in solving multivariate time series classification*". In this section, the results and applicability of DTW-DA-Net and ShapeDTW-1NN are assessed.

Section 4.3 evaluates the second objective "*To propose an incorporation of DTW with DA-Net; and compare to DA-Net"*. Performance of DTW-DA-Net is assessed in Section 4.3.1 and comparison with DA-Net is discussed in Section 4.3.2.

Section 4.4 tests the third objective "*To evaluate and compare the performance of DTW-DA-Net with other state-of-the-art multivariate time series classification methods*". DTW-DA-Net's performance is compared to existing distance-based methods; ED-1NN, DTW-1NN, and their variants.

Section 4.5 summarizes the outcome of the results.

4.2 APPLICABILITY OF UNIVARIATE DISTANCE-BASED CLASSIFICATION METHODS IN SOLVING MULTIVARIATE TIME SERIES CLASSIFICATION

There are many distance-based classification methods performed on TSC, but many are on univariate time series data. Our first objective is to assess if the recognized distance-based method on UTSC can be extended to use on MTSC with good performance.

Ruiz et al. (2021) suggests that DTW-D should be the benchmark for comparison. DTW-1NN-D by Chen et al. (2013) was one of the recognized methods that was used in the UEA multivariate dataset. There are many variants of DTW. Zhao & Itti, (2017) showed that NN-shape DTW was able to perform well with univariate data where it is better than DTW on many datasets and improved classification accuracy. In this dissertation, we use ShapeDTW-1NN to evaluate if we can extend the use of the DTW variant from univariate to multivariate application.

One of the most well-known distance-based methods for UTSC is DTW. We have selected DTW to incorporate into one of the recent well-performing methods, DA-Net for evaluation. DTW-DA-Net assesses the extensibility of DTW into a new algorithm and ShapeDTW-1NN's approach is to assess the extensibility of DTW variant from univariate to multivariate TSC.

Table 4.1 of performance comparison showed that ShapeDTW and DTW-DA-Net can perform on the four types of multivariate time series data. DTW-DA-Net has four wins whereas ShapeDTW has three wins. Performance comparison shows the difference of results in percentage for ShapeDTW-1NN to DTW-DA-Net and vice versa. DTW-DA-Net has only one win more than ShapeDTW, but by comparing the performance in percentage, the magnitude of improvement is much greater in DTW-DA-Net.

HMD and SWJ using DTW-DA-Net have 158.70%% and 150.38% improvement compared to only 12.85% in SRS2 which is the highest improvement using ShapeDTW. As HMD and SWJ have more classes compared to SRS2 with only two classes, DTW-DA-Net with the more complex self-learning algorithm may work better. For AF which only has a small train and test dataset of 15 each, there is no difference in performance where both methods have an accuracy of 0.267.

For electrical biosignals, DTW-DA-Net performs better with 2 wins compared to ShapeDTW-1NN where 1 win is of very low margin difference. DTW-DA-Net performs better on HAR datasets which has more than 2 classes to distinguish.

ShapeDTW also performs better on the HB dataset which has only two classes. This performance for ShapeDTW is better for motion data which has more than two classes but only with a negligible margin of 0.7.

The performance comparison is shown in Figure 4.1 where overall DTW-DA-Net performance is better than ShapeDTW-1NN with an average performance improvement of 40.39% (Table 4.1). The effect of noise in the time series may have more impact on ShapeDTW-1-NN resulting in poorer results.

Table 4.1 Performance comparison of ShapeDTW-1NN and DTW-DA-Net with optimal performance in bold

Type	**Dataset**	**Accuracy**		**Performance comparison (%)**	
		ShapeDTW-1NN	**DTW-DA-Net**	**ShapeDTW-1NN/ DTW-DA-Net**	**DTW-DA-Net/ ShapeDTW-1NN**
Electrical Biosignals (ECG/ EEG/ MEG)	AF	0.267	0.267	0.00	0.00
	HMD	0.230	**0.595**	-61.34	**158.70**
	SRS2	**0.527**	0.467	**12.85**	-11.39
	SWJ	0.133	**0.333**	-60.06	**150.38**
Accelerometer/ gyroscope (HAR)	BM	0.750	**0.975**	-23.08	**30.00**
	NA	0.867	**0.917**	-5.41	**5.72**
Coordinates (Motion)	AWR	**0.983**	0.976	**0.72**	-0.71
Audio (AS)	HB	**0.712**	0.644	**10.56**	-9.55
Average		0.56	0.65	-15.72	40.39
Win		3	4	3	4

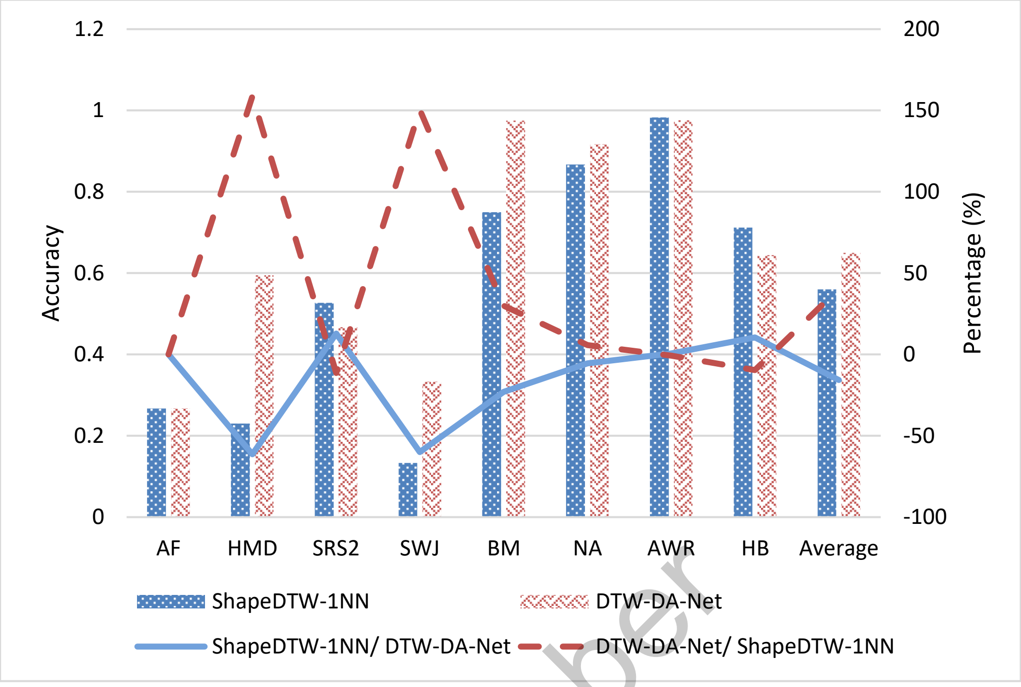
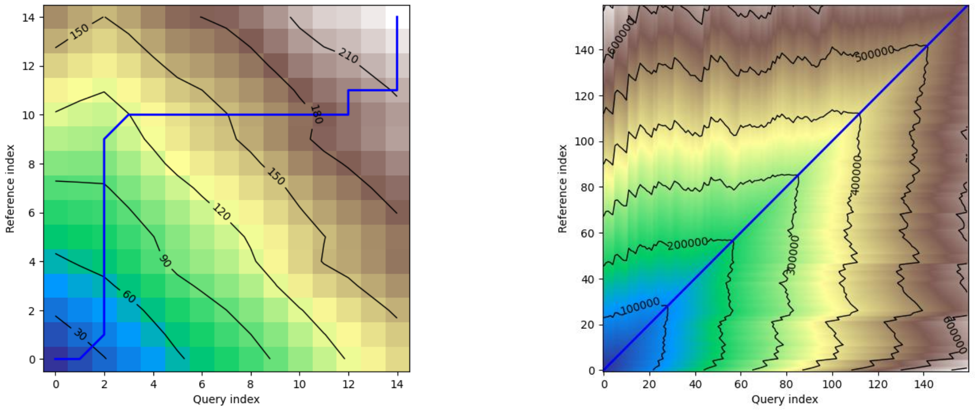


Figure 4.1 Performance comparison of ShapeDTW-1NN and DTW-DA-Net

In summary, DTW and ShapeDTW used for UTSC can be extended to use in MTSC; whether it is the incorporation of basic DTW with newer algorithms or addition of variant DTW to existing method. ShapeDTW-1NN generally performs better on datasets with 2 classes only and DTW-DA-Net performs better in datasets with more than two classes.

Figure 4.2 shows the warping path of the multivariate time series as one common path which is similar to the univariate time series.



AF

HMD

To be continued…

…continuation

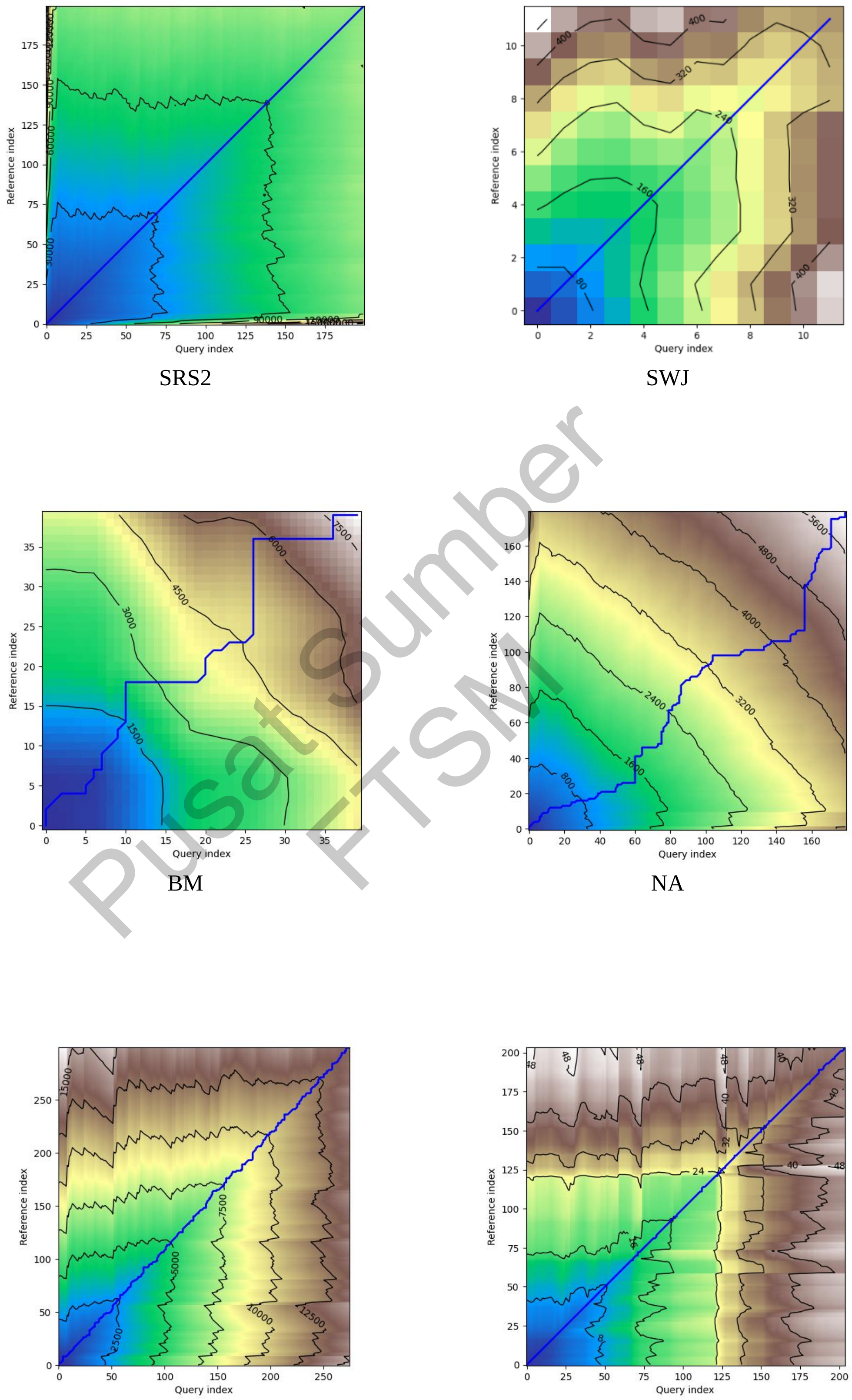


Figure 4.2 Warping path of density plot in the eight multivariate time series dataset

4.3 EVALUATION OF INCORPORATION OF DTW INTO DA-NET

4.3.1 Evaluation of DTW-DA-Net

Table 4.2 and Figure 4.3 shows the performance of DTW-DA-Net on the eight datasets. DTW-DA-Net performs very well for BM, NA, and AWR data, performs moderately for HB and HMD data, and does not perform well for AF, SRS2, and SWJ data.

The F1-score, precision, and recall reflect the accuracy of the method on the datasets. High accuracy of above 0.9 is seen on datasets with more than four classes but there is no consistent pattern in low and moderate performance in datasets with four or fewer classes. The best performers are observed on datasets with shorter lengths of below 150.

Table 4.2 Evaluation metrics on the eight datasets using DTW-DA-Net

Type	Dataset	Accuracy	F1 score	Precision	Recall
Electrical Biosignals (ECG/ EEG/ MEG)	AF	0.267	0.221	0.192	0.267
	HMD	0.595	0.563	0.572	0.583
	SRS2	0.467	0.467	0.467	0.467
	SWJ	0.333	0.274	0.259	0.333
Accelerometer/ gyroscope (HAR)	BM	**0.975**	**0.975**	**0.977**	**0.975**
	NA	**0.917**	**0.917**	**0.919**	**0.917**
Coordinates (Motion)	AWR	**0.977**	**0.977**	**0.979**	**0.977**
Audio (AS)	HB	0.644	0.597	0.596	0.613
Average		0.647	0.624	0.620	0.641

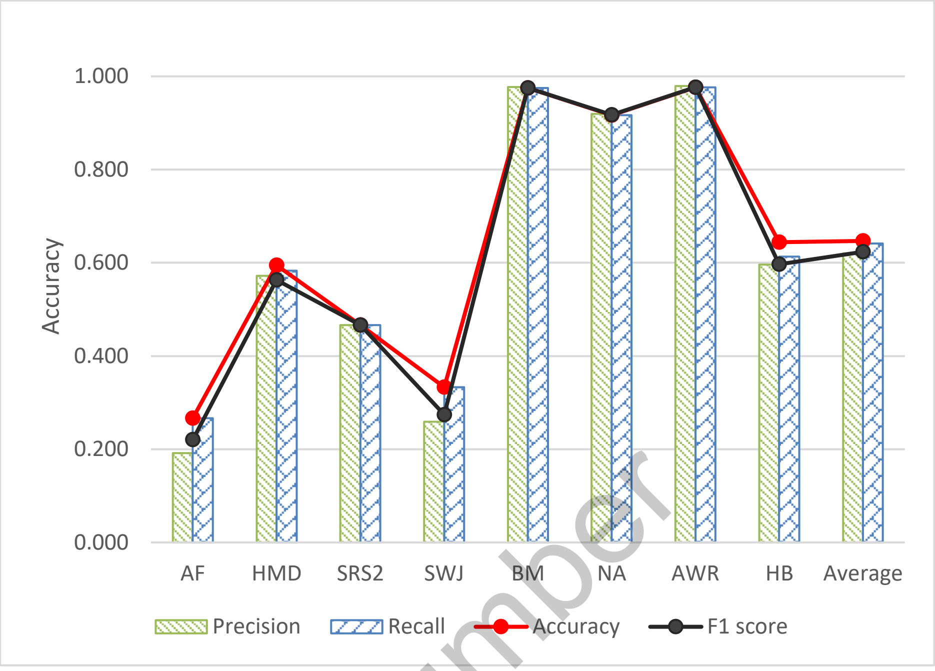
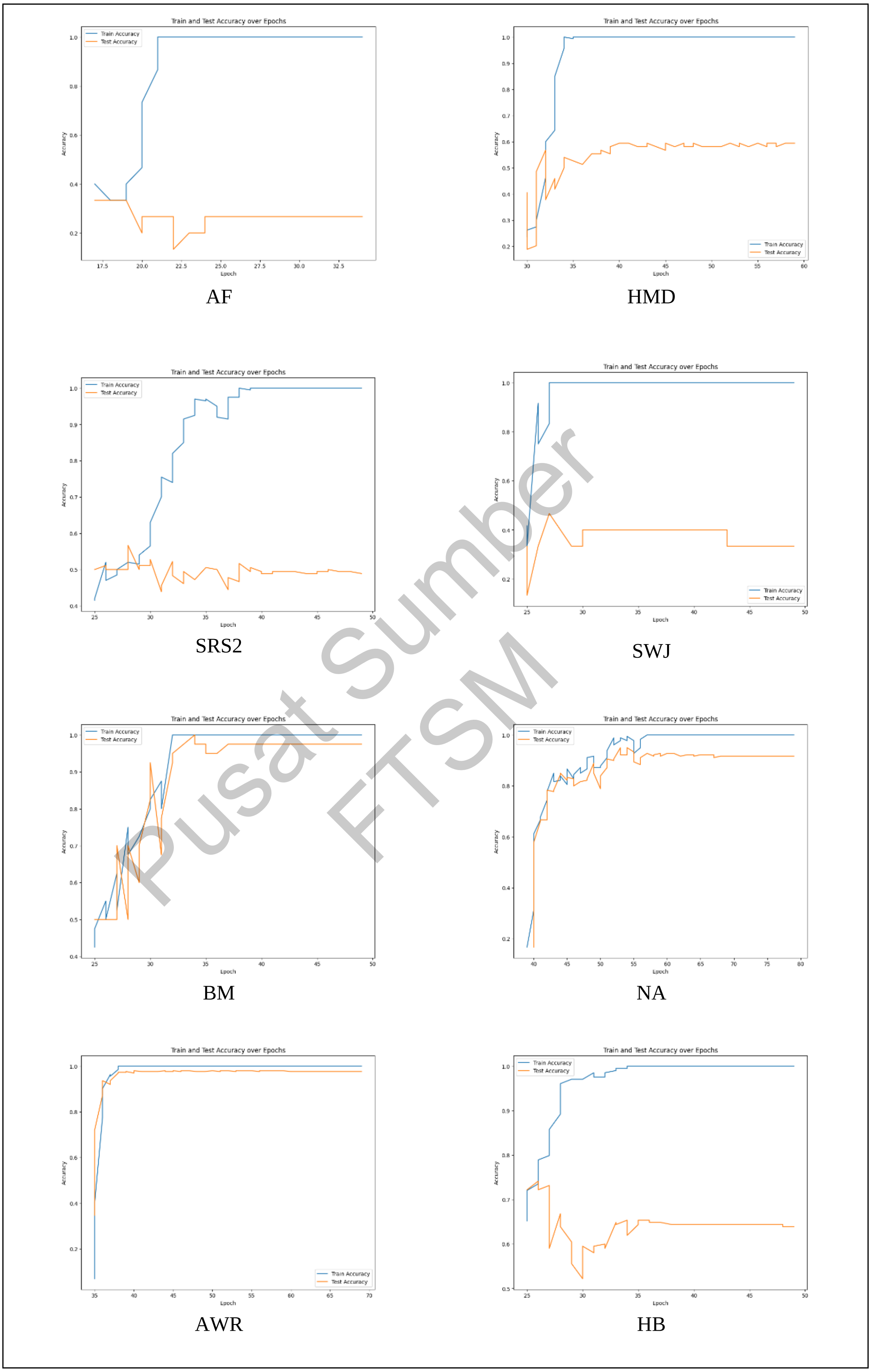


Figure 4.3 Visualization of evaluation metrics on the 8 datasets using DTW-DA-Net

Figure 4.4 shows the accuracy plot for DTW-DA-Net. BM, NA, and AWR show the merging of accuracy whereas there is a gap in the training and test accuracy curve for the other datasets. The large gap between the train and test set of AF, HMD, SRS2, SWJ, and HB suggests there may be overfitting. This may be due to the pattern of the data or type of algorithm.



AF

HMD

SRS2

SWJ

BM

NA

AWR

HB

Figure 4.4 DTW-DA-Net model accuracy plot. Train set in blue line and test set in orange line.

Figure 4.5 shows the time taken to train each dataset which ranges from 7 minutes to 143 minutes depending on data size. The longest time of 143 minutes is for the AWR dataset which is of coordinate type, with 275 train data, 300 train data, 9 dimensions, a length of 144, and 25 classes. In real-life situations, the dataset will be much larger than this and thus be a limitation to performing this method.

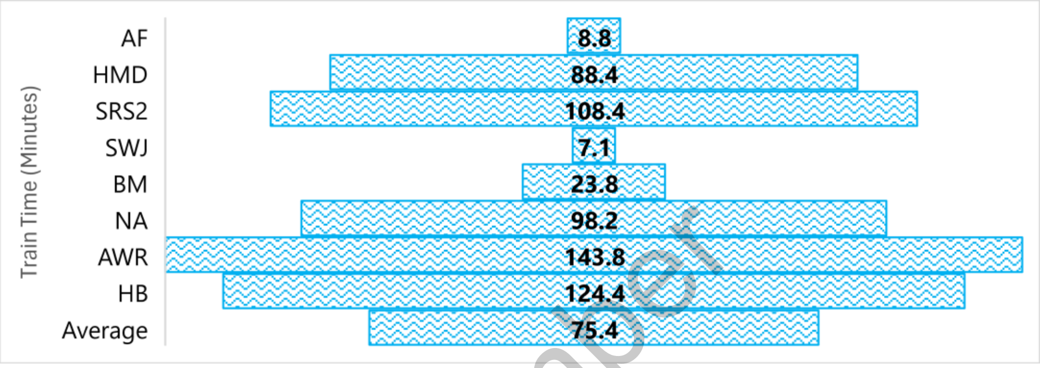


Figure 4.5 DTW-DA-Net train time for each dataset in minutes.

4.3.2 Comparison of DTW with DTW-DA-Net

DA-Net produced promising results on MTSC (Chen et al. 2022). The writers have proposed to look into DTW for dynamic window splitting in the future. In this dissertation, we look into the effect of DTW on DA-Net.

Table 4.3 of the performance comparison between DA-Net and DTW-DA-Net showed similar outcomes with both methods having four wins each. DA-Net has an average accuracy of 0.64 and DTW-DA-Net has an average accuracy of 0.65. The performance comparison is presented as a percentage improvement in DA-Net relative to DTW-DA-Net and vice versa. The highest increase in performance is by DTW-DA-Net on the HMD dataset with 71.47% improvement. BM, NA, and HB have slight improvements of 5.41%, 4.52%, and 2.88% respectively.

DA-Net still outperforms DTW-DA-generally for AF, SRS2 SWJ and AWR datasets; whereas DTW-DA-Net performs better for BM, NA, and HB. Coordinates and audio data types have only one dataset each and cannot be used to generalize.

Table 4.3 Performance comparison of DA-Net and DTW-DA-Net with optimal performance in bold

Type	Dataset	Accuracy		Performance Comparison (%)	
		DA-Net	DTW-DA-Net	DA-Net/DTW-DA-Net	DTW-DA-Net/DA-Net
Electrical Biosignals (ECG/ EEG/ MEG)	AF	**0.414**	0.267	**55.06**	-35.51
	HMD	0.347	**0.595**	-41.68	**71.47**
	SRS2	**0.561**	0.467	**20.13**	-16.76
	SWJ	**0.400**	0.333	**20.12**	-16.75
Accelerometer/ gyroscope (HAR)	BM	0.925	**0.975**	-5.13	**5.41**
	NA	0.877	**0.917**	-4.32	**4.52**
Coordinates (Motion)	AWR	**0.980**	0.976	**0.41**	-0.41
Audio (AS)	HB	0.626	**0.644**	-2.8	**2.88**
Average		0.64	0.65	**5.22**	1.86
Win		4	4	4	4

Figure 4.6 displays the similarities in accuracy in most of the datasets where the performance of both methods is comparable, except for HMD with a large difference.

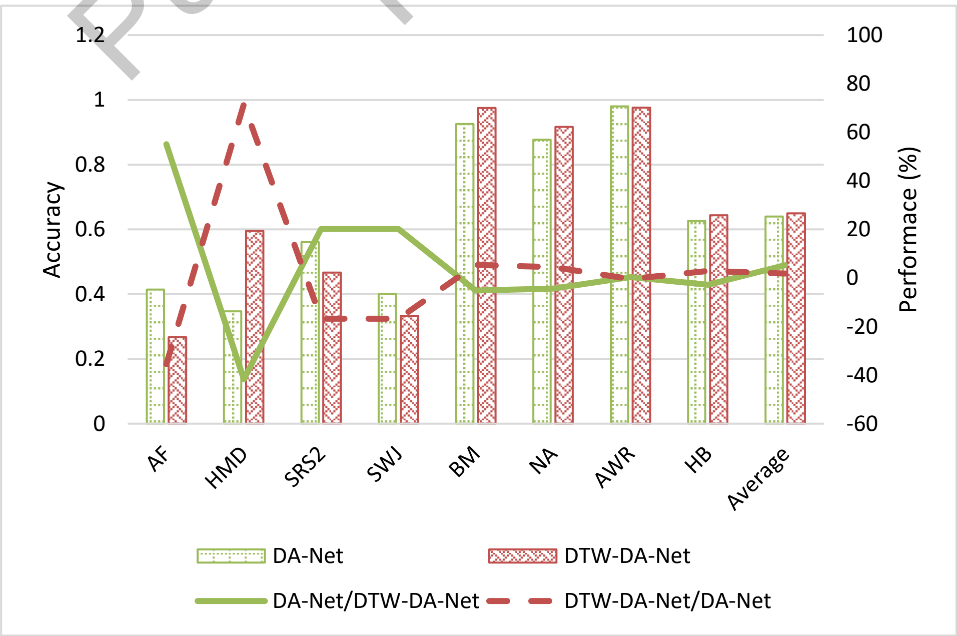


Figure 4.6 Performance comparison of DA-Net and DTW-DA-Net

In summary, the addition of DTW to DA-Net has a different outcome on different datasets and is not generalized to the data type. Both classifiers are comparable with DA-Net performing slightly better.

4.4 COMPARISON OF DTW-DA-NET WITH OTHER MULTIVARIATE TIME SERIES CLASSIFICATION METHODS

Table 4.4 shows the comparison to evaluate the performance of incorporation of DTW into DA-Net with well-performing distance-based methods which are ED-1NN, DTW-1NN, their variants, (Chen et al. 2013) and the recent DA-Net (Chen et al. 2022).

DTW -DA-Net achieved two wins, one for the HMD dataset with an accuracy of 0.595 and one for the NA dataset with 0.917 accuracy. ShapeDTW-1NN on the other hand did not manage to outperform any of the existing methods. BM and AWR have achieved very good results before and both methods in this dissertation perform poorer than the existing methods.

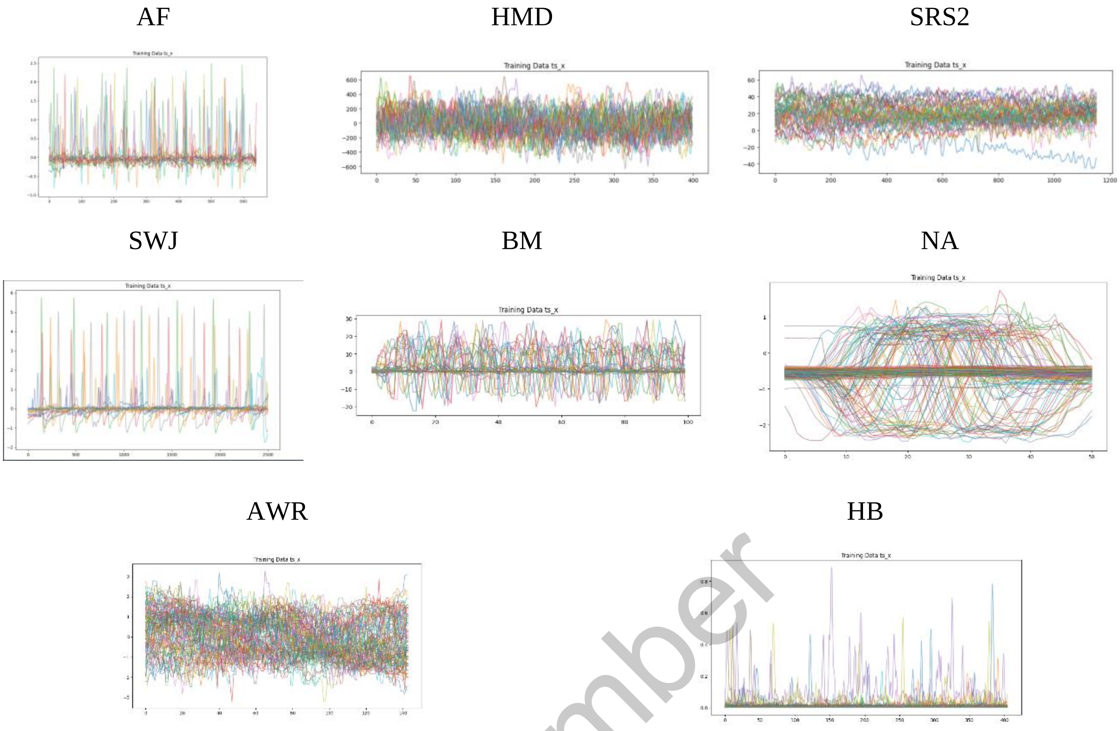
The highest average accuracy of 0.65 is DTW-DA-net, followed by DA-Net with 0.64 accuracy. However, DTW-DA-Net has 2 wins compared to DA-Net with three wins. These newer methods have slightly better overall performance compared to the existing DTW-1NN-I and DTW-1NN-I (norm) with average 0.62 accuracy and one win each.

The overall comparison showed that DA-Net performs well on AF, SRS2, and SWJ (electrical biosignals). DTW-DA-Net is suitable to perform best for NA and second best for BM which suggests beneficial in accelerometer/gyroscope dataset. DTW-DA-Net is unable to beat the existing DTW-1NN variant for AWR and HB.

One point to highlight with DTW-DA-Net is the improved performance in the NA dataset compared to the other methods. Figure 4.7 shows that NA has an unpredictable pattern compared to the other MTS data. The addition of DTW may be beneficial for irregular time series datasets.

Table 4.4 Comparison of accuracy in different methodologies with optimal performance in bold

	Accuracy								
Dataset	**ED-1NN**	**DTW-1NN-D**	**ED-1NN(norm)**	**DTW-1NN-I**	**DTW-1NN-I (norm)**	**DTW-1NN-D (norm)**	**DA-Net**	**ShapeDTW-1NN**	**DTW-DA-Net**
AF	0.267	0.200	0.247	0.267	0.267	0.220	**0.414**	0.267	0.267
HMD	0.279	0.231	0.278	0.306	0.303	0.231	0.347	0.230	**0.595**
SRS2	0.483	0.539	0.483	0.533	0.533	0.539	**0.561**	0.527	0.467
SWJ	0.200	0.200	0.200	0.333	0.333	0.200	**0.400**	0.133	0.333
BM	0.675	0.975	0.676	**1.000**	**1.000**	0.975	0.925	0.750	0.975
NA	0.860	0.883	0.850	0.850	0.85	0.883	0.877	0.867	**0.917**
AWR	0.970	**0.987**	0.970	0.980	0.980	**0.987**	0.980	0.983	0.976
HB	0.620	**0.717**	0.619	0.659	0.658	**0.717**	0.626	0.712	0.644
Average	0.54	0.47	0.54	0.62	0.62	0.59	0.64	0.56	0.65
Win	0	0	0	1	1	0	3	0	2



AF HMD SRS2

SWJ BM NA

AWR HB

Figure 4.7 Summary of visualization of eight datasets.

Figure 4.8 shows a critical difference diagram for the two new methods of ShapeDTW-1NN and DTW-DA-Net compared to six existing distance-based methods and the recent DA-Net. All classifiers are compared using the Friedman test and ranked based on the Wilcoxon test with the smallest number representing the best performance. We test the null hypothesis that there is no difference between the accuracy of all the methods at $\alpha = 0.05$ significance level. Null hypotheses is rejected if $p \leq \alpha$ and fail to reject if $p \geq \alpha$ We fail to reject the null hypothesis The horizontal thick line across indicates that there is no difference in performance for the methods. The existing DA-Net still performs best compared to the other methods for these eight datasets.

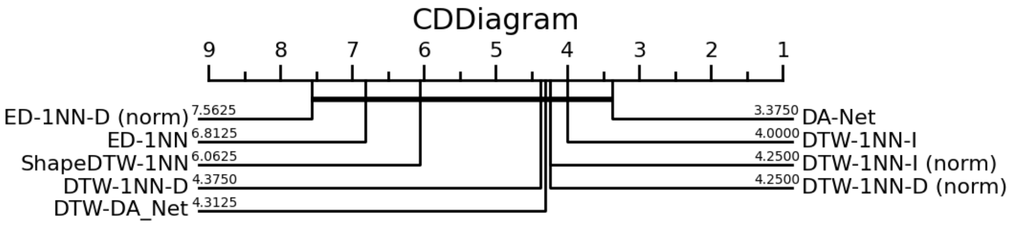


Figure 4.8 Critical difference diagram for Shape-DTW and DTW-DA-Net compared to seven existing methods

Table 4.5 shows example of the *p*-value of different algorithms with relative to DTW-1NN-D. These values will be compared with an adjusted α value by Holm-Bonferroni correction which will not be covered in the scope of this dissertation

Table 4.5 Performance relative to benchmark DTW-1NN-D

Algorithm	**P-value**
ShapeDTW-1NN	0.109
ED-1NN-D(norm)	0.233
ED-1NN	0.292
DTW-1NN-D(norm)	0.317
DTW-1NN-I	0.461
DTW-1NN-I(norm)	0.461
DA-Net	0.461
DTW-DA-Net	0.528

Figure 4.9 shows a box plot of the eight datasets based on the results of Table 4.4. BM, NA (accelerometer/gyroscope data), and AWR (coordinates) data are the best performers with most algorithms. BM has the widest distribution of accuracy indicating nonconsistency in methods and difficulty in generating accurate results. AWR has the best performance as all the methods give good results with small variances. NA also performs well and is consistent with all algorithms. HB and SRS2 showed intermediate performance outcomes with most algorithms. HB is an audio dataset whereas SRS2 is electrical biosignals. These two datasets have a similar number of train and test datasets with the least class of two but differ in dimension and length. The similarities may lead to similar performance of different classes. AF, HMD, and SWJ which are electrical biosignals are the lowest-performing datasets across all algorithms.

AF and HMD always had poor results, but DA-Net and DTW-DA-Net were able to produce a major improvement in results as can be seen in the outlier 0.414 for AF with DA-Net, and 0.595 for HMD with DTW-DA-Net. By comparing HMD within the electrical biosignal dataset, it has the highest number of classes, 4, and the largest dimension, 10. The incorporation of DTW into DA-Net has affected the interpretation of data for better outcomes. AF has consistently produced results below 0.267 but DA-Net was able to improve it to 0.414 compared to DTW-DA-Net which is still 0267. AF is made of only 15 train and test datasets. It may be possible that the incorporation of DTW into the algorithm with a small amount of data has produced overfitting in the deep learning network.

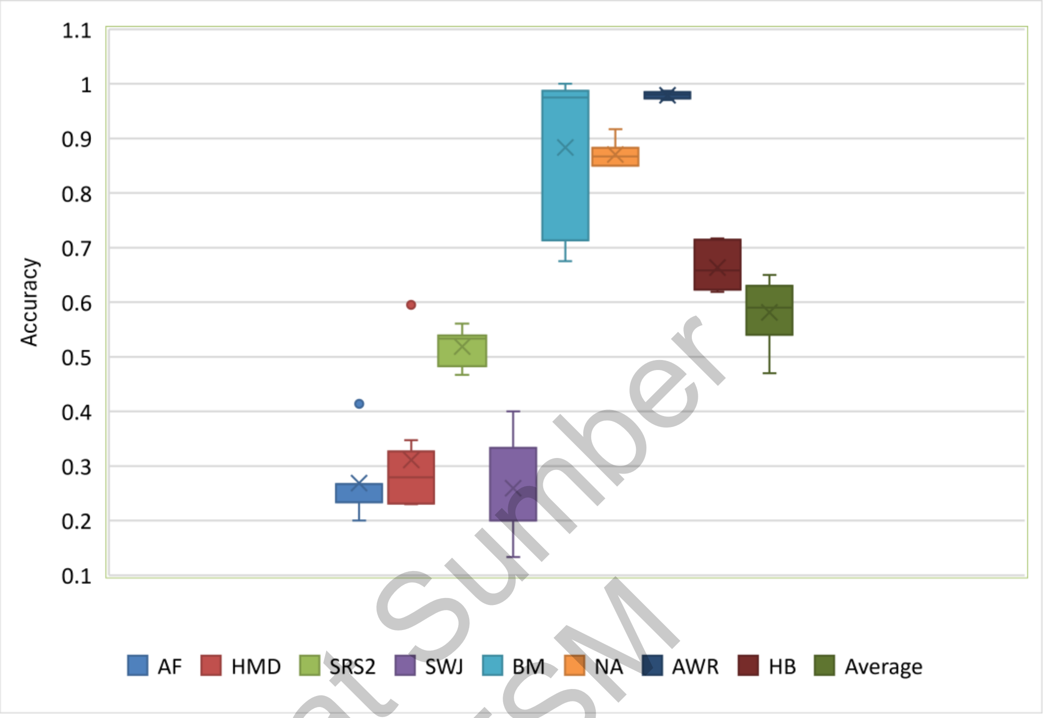


Figure 4.9 Box plots of differences in accuracies of different methods on different datasets.

In short, DTW-DA-Net can produce better results in two out of eight of the existing datasets. There is still no one-size-fits-all algorithm for MTSC. The selection of the algorithm may still be dependent on the data type.

4.5 DISCUSSION

Differences in performance between the methods can be due to many factors such as differences in properties of datasets, methods, and the suitability or applicability of the methods to the dataset.

Multivariate time series data can be in different number of dimensions, lengths, and classes. Complicated data with a larger dimension, longer sequence, and more classes are more difficult to classify. A different number of training data also affects the machine algorithm's training. Training a dataset that is too small will not be sufficient to teach the algorithm to perform properly. It is a challenge to extend UTSC into MTSC as the structure of data for both has a major difference. Classification for multivariate time series needs to consider many aspects especially on the local, global and spatial features. The relation and correlation between these features can be very complex. No one algorithm can suit all datasets. In this research, DTW-DA-Net has the benefit of the NA dataset which performs better than other existing methods with an accuracy of 0.917 and the accuracy curve indicates no overfitting.

Methods like Shape-DTW with 1-NN use a nearest-neighbor approach and are beneficial for data where shape discrimination is important. DTW-DA-Net is a deep learning method with a neural network approach and can manage more complex datasets.

4.6 CHAPTER SUMMARY

This chapter presents the outcome of the research questions and objectives. Section 4.2 shows the outcomes of the applicability of UTSC methods to MTSC in response to the first research question and objective.

Section 4.3 regarding the second research question and objective showed that DTW-DA-Net performs similarly to DA-Net with four wins each. The accuracy for DTW-DA-Net is 0.65 and DA-Net is 0.64.

Evaluation of the third research question and objective in Section 4.4 showed that DTW-DA-Net performed better in two out of eight datasets when compared to other distance-based methods.

CHAPTER V

CONCLUSION AND FUTURE WORK

5.1 INTRODUCTION

This chapter concludes the discussions and findings of the proposed method, explains the limitations, and suggests future work. This chapter of three sections consists of Section 5.2 which is the research summary, and Section 5.3 which is the limitation and future research direction.

5.2 RESEARCH SUMMARY

There are many established methods in UTSC. However, for MTSC, there are still limited methods of producing consistent results with different types of datasets. In reality, there are more MTSC problems than UTSC problems and it is still a challenge to determine a generalized method for MTSC problems. This research approaches this challenge by reviewing available established methods and recent well-performing methods to produce a new proposed algorithm by expanding the distance-based method from UTSC to MTSC.

To address the first research question "*Are distance-based methods such as DTW and ShapeDTW performed on univariate time series analysis applicable to multivariate time series?*" and the first objective "*To assess the applicability of univariate distance-based classification methods in solving multivariate time series classification*.", DTW is incorporated into DA-Net and its variant, ShapeDTW is incorporated into 1-NN. These methods showed that DTW and its variant ShapeDTW can be used for MTSC. Comparison of results showed that incorporation of DTW into recent methods produced better outcomes than incorporating variant DTW into traditional algorithms.

DTW with 1-NN is an established method and is still recommended as the benchmark for MTSC. DA-Net is a recent algorithm that produced improved performance compared to existing MTSC methods. This research proposed to incorporate established DTW with recent DA-Net to answer the second research question "*Does incorporation of DTW to DA-Net affect performance of multivariate time series classification?*" and address the second objective "*To propose an incorporation of DTW with DA-Net; and compare to DA-Net*". Comparison of DTW-DA-Net with DA-Net showed different suitability for different types of data. DTW-DA-Net performs best with HAR and AS data. DTW-DA-Net does not perform well on most ECG data.

DTW-DA-Net results are compared with existing distance-based methods to answer the research question "*How is the performance of DTW-DA-Net compared with other state-of-the-art multivariate time series classification methods?*" and address the objective "*To evaluate and compare the performance of DTW-DA-Net with other state-of-the-art multivariate time series classification methods.*". There is no significant difference in all the algorithms tested. However, DTW-DA-Net produced the best results in two out of eight of the datasets when compared to existing distance-based methods. DTW-DA-Net achieved the best accuracy of 0.917 in the NA dataset.

5.3 LIMITATION AND FUTURE RESEARCH DIRECTION

MTSC is a complex problem and there is room for much improvement. The limitations in this dissertation can serve as opportunities for further consideration in future works

1. This dissertation used dependent DTW. Future studies can be performed on independent DTW as there are still disputes on which DTW performs better for which kind of dataset.

2. DTW-DA-Net takes a long time to run on large datasets and may not be suitable for practical use. It is best to evaluate ways for DTW to perform faster such as implementing boundaries or limitations to improve speed for use on large datasets.

3. The evaluation of DTW-DA-Net and ShapeDTW-1NN with existing methods showed that there is an improvement in performance for DTW-DA-Net but no improvement for ShapeDTW-1NN. It is recommended to look into future methods that incorporate standard DTW into newer algorithms instead of variants of DTW into traditional methods.

4. The most improvement is seen in the HMD dataset with an accuracy of 0.595 compared to the previous best accuracy of 0.347 by DA-Net. HMD has the highest number of classes and largest dimension in the group of four electrical biosignals. There is also improvement in NA which is one out of two accelerator/gyroscope datasets. This data also has a higher dimension and class compared to the other datasets in the same group. As there are only four data on electrical biosignals and two on accelerator/gyroscope in this dissertation, there may be a potential benefit to look into further analysis in this data type with larger dimension or class to assess if the incorporation of DTW is of significance.

5. The accuracy plot shows that there is a gap between the train and test set for AF, HMD, SRS2, SWJ, and HB. There may be overfitting. It is suggested to perform the proposed method on a larger train dataset for better accuracy.

REFERENCES

Abbasi, M. & Saeedi, P. 2023. Enhancing Multivariate Time Series Classifiers through Self-Attention and Relative Positioning Infusion. IEEE

Altuwaijri, G.A., Muhammad, G., Altaheri, H. & Alsulaiman, M. 2022. A Multi-Branch Convolutional Neural Network with Squeeze-and-Excitation Attention Blocks for EEG-Based Motor Imagery Signals Classification. *Diagnostics* 12(4).

Bagnall, A., Keogh, E., Lines, J., Bostrom, A., Large, J. & Middlehurst, M. (n.d.). Time Series Classification. https://www.timeseriesclassification.com/index.php [26 May 2024].

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.

Bellman, R.E. 1957. *Dynamic Programming*. USA: Princeton Universoty Press.

Bignoumba, N., Mellouli, N. & Yahia, S. Ben. 2024. A new efficient ALignment-driven Neural Network for Mortality Prediction from Irregular Multivariate Time Series data[Formula presented]. *Expert Systems with Applications* 238.

Bińkowski, M., Marti, G. & Donnat, P. 2017. Autoregressive Convolutional Neural Networks for Asynchronous Time Series.

Bostrom, A. & Bagnall, A. 2017. Binary Shapelet Transform for Multiclass Time Series Classification. , pp. 24–46.

Cao, D. & Liu, J. 2016. Research on dynamic time warping multivariate time series similarity matching based on shape feature and inclination angle. *Journal of Cloud Computing* 5(1): 11.

Chen, R., Yan, X., Wang, S. & Xiao, G. 2022a. DA-Net: Dual-attention network for multivariate time series classification. *Information Sciences* 610: 472–487.

Chen, Y., Hu, B., Keogh, E. & Batista, G.E.A.P.A. 2013. DTW-D: Time series semi-supervised learning from a single example. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 383–391.

Chen, Z., Liu, Y., Zhu, J., Zhang, Y., Jin, R., He, X., Tao, J. & Chen, L. 2021. Time-frequency deep metric learning for multivariate time series classification. *Neurocomputing* 462: 221–237.

Dau, H.A., Silva, D.F., Petitjean, F., Forestier, G., Bagnall, A., Mueen, A. & Keogh, E. 2018. Optimizing dynamic time warping’s window width for time series data mining applications. *Data Mining and Knowledge Discovery* 32(4): 1074–1120.

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.

Deng, H., Runger, G., Tuv, E. & Vladimir, M. 2013a. A time series forest for classification and feature extraction. *Information Sciences* 239: 142–153.

Du, M., Wei, Y., Zheng, X. & Ji, C. 2023. Multi-feature based network for multivariate time series classification. *Information Sciences* 639.

Escudero-Arnanz, Ó., Marques, A.G., Soguero-Ruiz, C., Mora-Jiménez, I. & Robles, G. 2023. dtwParallel: A Python package to efficiently compute dynamic time warping between time series. *SoftwareX* 22.

Ghouaiel, N., Marteau, P.F. & Dupont, M. 2017. Continuous pattern detection and recognition in stream - a benchmark for online gesture recognition. *International Journal of Applied Pattern Recognition* 4(2): 146.

Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.Ch., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.-K. & Stanley, H.E. 2000. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation* 101(23).

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J. & Chen, T. 2018. Recent advances in convolutional neural networks. *Pattern Recognition* 77: 354–377.

Hao, Y. & Cao, H. 2020. A New Attention Mechanism to Classify Multivariate Time Series.

Hao, Y., Cao, H., Mueen, A. & Brahma, S. 2021. Identify Significant Phenomenon-Specific Variables for Multivariate Time Series. *IEEE Transactions on Knowledge and Data Engineering* 33(3): 1019–1031.

Herrmann, M., Tan, C.W. & Webb, G.I. 2023. Parameterizing the cost function of dynamic time warping with application to time series classification. *Data Mining and Knowledge Discovery* 37(5): 2024–2045.

Herrmann, M. & Webb, G.I. 2023. Amercing: An intuitive and effective constraint for dynamic time warping. *Pattern Recognition* 137: 109333.

Hsu, C.-J., Huang, K.-S., Yang, C.-B. & Guo, Y.-P. 2015. Flexible Dynamic Time Warping for Time Series Classification. *Procedia Computer Science* 51: 2838–2842.

Hu, J., Shen, L., Albanie, S., Sun, G. & Wu, E. 2017. Squeeze-and-Excitation Networks.

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L. & Muller, P.A. 2019. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery* 33(4): 917–963.

Iwana, B.K., Frinken, V. & Uchida, S. 2020. DTW-NN: A novel neural network for time series recognition using dynamic alignment between inputs and weights. *Knowledge-Based Systems* 188.

Jeong, Y.-S., Jeong, M.K. & Omitaomu, O.A. 2011. Weighted dynamic time warping for time series classification. *Pattern Recognition* 44(9): 2231–2240.

Karim, F., Majumdar, S., Darabi, H. & Harford, S. 2018. Multivariate LSTM-FCNs for Time Series Classification.

Keogh, E.J. & Pazzani, M.J. 2001. Derivative Dynamic Time Warping. *Proceedings of the 2001 SIAM International Conference on Data Mining*, pp. 1–11.

Lian, Z., Liu, B. & Tao, J. 2021. CTNet: Conversational Transformer Network for Emotion Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29: 985–1000.

Liang, Z. & Wang, H. 2021. Efficient class-specific shapelets learning for interpretable time series classification. *Information Sciences* 570: 428–450.

Lines, J. & Bagnall, A. 2015. Time series classification with ensembles of elastic distance measures. *Data Mining and Knowledge Discovery* 29(3): 565–592.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. & Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002.

Lubba, C.H., Sethi, S.S., Knaute, P., Schultz, S.R., Fulcher, B.D. & Jones, N.S. 2019. catch22: CAnonical Time-series CHaracteristics. *Data Mining and Knowledge Discovery* 33(6): 1821–1852.

Lucas, B., Shifaz, A., Pelletier, C., O’Neill, L., Zaidi, N., Goethals, B., Petitjean, F. & Webb, G.I. 2019a. Proximity Forest: an effective and scalable distance-based classifier for time series. *Data Mining and Knowledge Discovery* 33(3): 607–635.

Lucas, B., Shifaz, A., Pelletier, C., O’Neill, L., Zaidi, N., Goethals, B., Petitjean, F. & Webb, G.I. 2019b. Proximity Forest: an effective and scalable distance-based classifier for time series. *Data Mining and Knowledge Discovery* 33(3): 607–635.

MacKay, D.J.C. 1992. Bayesian Interpolation. *Neural Computation* 4(3): 415–447.

Melhem, S., Al-Aiad, A. & Al-Ayyad, M.S. 2021. Patient care classification using machine learning techniques. *2021 12th International Conference on Information and Communication Systems (ICICS)*, pp. 57–62.

Middlehurst, M. & Bagnall, A. 2022. The FreshPRINCE: A Simple Transformation Based Pipeline Time Series Classifier. , pp. 150–161.

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.

Middlehurst, M., Schäfer, P. & Bagnall, A. 2023. Bake off redux: a review and experimental evaluation of recent time series classification algorithms. *Data Mining and Knowledge Discovery* 1-74.

Mohammadi Foumani, N., Miller, L., Tan, C.W., Webb, G.I., Forestier, G. & Salehi, M. 2024. Deep Learning for Time Series Classification and Extrinsic Regression: A Current Survey. *ACM Computing Surveys* 56(9): 1–45.

Oastler, G. & Lines, J. 2019. A Significantly Faster Elastic-Ensemble for Time-Series Classification. , pp. 446–453.

Orsenigo, C. & Vercellis, C. 2010. Combining discrete SVM and fixed cardinality warping distances for multivariate time series classification. *Pattern Recognition* 43(11): 3787–3794.

Qiu, S., Zhao, H., Jiang, N., Wang, Z., Liu, L., An, Y., Zhao, H., Miao, X., Liu, R. & Fortino, G. 2022. Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges. *Information Fusion* 80: 241–265.

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N. & Aigrain, S. 2013. Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371(1984): 20110550.

Ruiz, A.P., 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.

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.

Salvador, S. & Chan, P. 2007. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis* 11(5): 561–580.

Sammour, M., Ali Othman, Z., Mabrina Masbar Rus, A. & Mohamed, R. 2019. Modified Dynamic Time Warping for Hierarchical Clustering 9(5).

Schäfer, P. & Leser, U. 2017a. Fast and Accurate Time Series Classification with WEASEL. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 637–646.

Schäfer, P. & Leser, U. 2017b. Multivariate Time Series Classification with WEASEL+MUSE. 3rd ECML/PKDD workshop on AALTD

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.

Shifaz, A., Pelletier, C., Petitjean, F. & Webb, G.I. 2023. Elastic similarity and distance measures for multivariate time series. *Knowledge and Information Systems* 65(6): 2665–2698.

Shokoohi-Yekta, M., Hu, B., Jin, H., Wang, J. & Keogh, E. 2017a. Generalizing DTW to the multi-dimensional case requires an adaptive approach. *Data Mining and Knowledge Discovery* 31(1): 1–31.

Shokoohi-Yekta, M., Hu, B., Jin, H., Wang, J. & Keogh, E. 2017b. Generalizing DTW to the multi-dimensional case requires an adaptive approach. *Data Mining and Knowledge Discovery* 31(1): 1–31.

Sisodia, D. & Sisodia, D.S. 2018. Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science* 132: 1578–1585.

Tavenard Romain. 2021. An Introduction to Dynamic Time Warping. https://rtavenar.github.io/blog/dtw.html [31 May 2024].

Vapnik, V. & Golowich, S.E. (n.d.). Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing·.

Wang, J., Yang, C., Jiang, X. & Wu, J. 2023. WHEN: A Wavelet-DTW Hybrid Attention Network for Heterogeneous Time Series Analysis. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2361–2373.

Wang Jun, Balasubramanian Arvind, La Vega Luis Mojica, Green Jordan R, Samal Ashok & Prabhakaran Balakrishnan. 2013. Word Recognition from Continuous Articulatory Movement Time-Series Data using Symbolic Representations. *Proceedings of the 4th workshop on speech and language processing for assistive technologies*, pp. 119–127.

Wang, Z., Yan, W. & Oates, T. 2017. Time series classification from scratch with deep neural networks: A strong baseline. *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 1578–1585.

Xiao, Z., Xu, X., Xing, H., Luo, S., Dai, P. & Zhan, D. 2021. RTFN: A robust temporal feature network for time series classification. *Information Sciences* 571: 65–86.

Zerveas, G., Jayaraman, S., Patel, D., Bhamidipaty, A. & Eickhoff, C. 2021. A Transformer-based Framework for Multivariate Time Series Representation Learning. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2114–2124.

Zhang, C., Song, D., Chen, Y., Feng, X., Lumezanu, C., Cheng, W., Ni, J., Zong, B., Chen, H. & Chawla, N. V. 2019. A Deep Neural Network for Unsupervised

Anomaly Detection and Diagnosis in Multivariate Time Series Data. *Proceedings of the AAAI Conference on Artificial Intelligence* 33(01): 1409–1416.

Zhang, X., Gao, Y., Lin, J. & Lu, C.-T. 2020. TapNet: Multivariate Time Series Classification with Attentional Prototypical Network. *Proceedings of the AAAI Conference on Artificial Intelligence* 34(04): 6845–6852.

Zhao, J. & Itti, L. 2017. shapeDTW: shape Dynamic Time Warping.*Pattern Recognition* 74:171-184