Modelling Atmospheric Ozone Concentration Using Machine Learning Algorithms

By

Eman Said Al-Abri

A doctoral thesis

Submitted in partial fulfilment of requirements

for the award of

Doctor of Philosophy

Department of Computer Science

Loughborough University

November 2016

Supervisor: Prof. Eran Edirisinghe

Dr.Christian Dawson

Abstract

Air quality monitoring is one of several important tasks carried out in the area of environmental science and engineering. Accordingly, the development of air quality predictive models can be very useful as such models can provide early warnings of pollution levels increasing to unsatisfactory levels. The literature review conducted within the research context of this thesis revealed that only a limited number of widely used machine learning algorithms have been employed for the modelling of the concentrations of atmospheric gases such as ozone, nitrogen oxides etc. Despite this observation the research and technology area of machine learning has recently advanced significantly with the introduction of ensemble learning techniques, convolutional and deep neural networks etc. Given these observations the research presented in this thesis aims to investigate the effective use of ensemble learning algorithms with optimised algorithmic settings and the appropriate choice of base layer algorithms to create effective and efficient models for the prediction and forecasting of specifically, ground level ozone (O_3).

Three main research contributions have been made by this thesis in the application area of modelling O_3 concentrations. As the first contribution, the performance of several ensemble learning (Homogeneous and Heterogonous) algorithms were investigated and compared with all popular and widely used single base learning algorithms. The results have showed impressive prediction performance improvement obtainable by using meta learning (Bagging, Stacking, and Voting) algorithms. The performances of the three investigated meta learning algorithms were similar in nature giving an average 0.91 correlation coefficient, in prediction accuracy. Thus as a second contribution, the effective use of feature selection and parameter based optimisation was carried out in conjunction with the application of Multilayer Perceptron, Support Vector Machines, Random Forest and Bagging based learning techniques providing significant improvements in prediction accuracy. The third contribution of research presented in this thesis includes the univariate and multivariate forecasting of ozone concentrations based of optimised Ensemble Learning algorithms. The results reported supersedes the accuracy levels reported in forecasting Ozone concentration variations based on widely used, single base learning algorithms.

In summary the research conducted within this thesis bridges an existing research gap in big data analytics related to environment pollution modelling, prediction and forecasting where present research is largely limited to using standard learning algorithms such as Artificial Neural Networks and Support Vector Machines often available within popular commercial software packages.

Eman Al-Abri, October 2016

Dedicated to,

My Husband Fahmi Al-Zakwani and

My Kids Ghida, Qusai, and Jassar,

For their

Inspiration,

Sacrifices,

Love

Acknowledgment

First I would like to thank Allah (GOD), for his guidance and blessings in the completion of this research.

My biggest gratitude goes to my husband Fahmi Al Zakwani for his limitless support and encouragement through-out the years I have spent far from home. He always encouraged me to have positive and independent thinking. To my children Ghida, Qusai and Jassar from whom I built up courage and patience by merely thinking of them. I also would like to thank my sister in-law Maryam Al Zakwani and her husband Ali Al Riyami for their continuous support, love and caring for my children as one of their own for the past 3 years.

My sincere thanks goes to my family for always fueling me with positive encouragement and cheering.

Furthermore, I would like to express my deep thanks to my supervisors Professor Eran A Edirisinghe and Dr. Christian Dawson for all the help and supervision they provided that extremely assisted in the completion of this research. Also, my thanks goes to Dr. Amin for all the help and support he provided.

Moreover, my thanks goes to all of my friends for their love and support during the 3 years. special thanks to Asmaa AlMarhobi, Suaad AlSawafi, Muna AlRahbi and Shadha AlAmri for their supportive friendship and inspiration. Not to forget all the Omani families in Loughborough especially Al Batashi, Al Droushi and Al Harbi whom always were there to provide me with help and company whenever I needed.

In addition, my gratitude goes to Sohar University, Oman for providing me with the necessary data and information in order to initiate this research.

Finally, I would like to thank the Ministry of Manpower for providing the financial support through-out the research period.

Eman S. AlAbri, 4th October 2016

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ANN	Artificial Neural Network
BTX	Benzene , Toluene, and (o, m, p)-Xylene
CC	Correlation Coefficient
CO	Carbon monoxide
DOAS	Deferential Optical Absorption System
LWL	Locally Weighted Learning
MAE	Mean Absolute Error
ML	Machine Learning
MLR	Multiple Linear Regression
MLP	Multilayer Perceptron
M5R	Model Trees Rules
M5P	Model Trees Regression
NO	Nitrogen Monoxide
NO_2	Nitrogen Dioxide
NOx	anthropogenic nitrogen oxides
O_3	Ozone
PC	principle component
PCA	principle component analysis
PM10	Particulate Matter
R^2	coefficient of determination
RAE	Relative Absolute Error
REPTree	Reduced Error Pruning Tree
RF	Random Forest
RMSE	Root Mean Squared Error
RRSE	Root Relative Squared Error
SHW	Sohar Highway
SLR	Simple Linear Regression
SMOreg	Support Vector Machine for Regression

SO_2	Sulphur Dioxide
SU	Sohar University
SVM	Support Vector Machine
T	Temperature
TSF	Time Series analysis and Forecasting toolkit
VOCs	volatile organic compounds
WEKA	Waikato Environment for Knowledge Analysis toolkit
WD	Wind Direction
WS	Wind Speed

CHAPTER 1

An Overview

1.1. Introduction

In recent years, the environmental risks caused by exposure to ground level ozone (O_3) from both stationary and mobile sources have increased annually. Ozone is a transboundary air pollutant that can be formed by photochemical reactions between anthropogenic nitrogen oxides (NO_x) and volatile organic compounds (VOCs) in the presence of sunlight [1]. There are several sources which produce the particles that react to form ozone. Example of these sources are oil refining, printing, and motor vehicle exhaust. In addition exhaust emissions from motor vehicles are considered to be one of the major sources of pollution where it produces 70% of nitrogen oxides (main chemical to form ozone) and 50% of the organic chemicals (e.g. benzene, toluene and xylene)[2] that are normally present in the atmosphere of urban regional. When O_3 is formed, depending on meteorological conditions, it remains suspended in the lower atmosphere for hours to days and can endanger local and regional receptors. Several studies that analyse the effects of meteorological conditions on the formation and transport of O_3 have been listed in the work of [3]. Further, statistically significant relationships have been identified between elevated concentrations of O_3 and environmental risks [4]–[7].

The existing research on modelling atmospheric O_3 concentration for air pollution prediction makes use of widely employed traditional machine learning algorithms such as, Artificial Neural Networks (ANN) and Support Vector Machines (SVM). Although there are more advanced and novel data mining techniques, such as Ensemble Learning approaches, only few attempts have utilised them to predict the ozone concentration; these include the work of [8]–[10]. On the other hand, ensemble learning algorithms have been successfully used in statistical approaches and machine learning used in other application fields. Based on the literature review, the use of ensemble learning

techniques has been found to enhance the accuracy of the predictive models obtained compared to single base learning approaches such as ANN and SVM. However, these studies have also been limited to making use of a few selected ensemble classifiers namely Bagging, Stacking, Voting and Random forest, with no comparisons between the use of different techniques or investigations of scientific rigour. It is also known that the classification approach that works best is dependent on the data and the application domain. Hence a significant research gap exists in the investigation of the best prediction models for atmospheric ozone as no detailed and rigourous investigations have been carried out.

Therefore, the research proposed in Chapter 5 and Chapter 6 of this thesis aims to find the most accurate machine learning algorithms and models that can be used to predict ground level ozone concentrations, given a multitude of meteorological parameters and concentrations of gases that are known to create ozone by decomposition due to natural causes, e.g., NO_x. We describe such a prediction approach a *spatial prediction approach* as no time-dependent data, for e.g. concentration of ozone at given times of the day, is considered. Instead the time at which ozone concentration is recorded, is completely ignored and only a concentration value and the parameters that have known to cause such a value is considered for modelling. A comprehensive investigation was carried out comparing the performance of several machine learning techniques. In the research proposed, multiple predictive models were built using 13 single based algorithms and four ensemble learning algorithms using the WEKA (Waikato Environment for Knowledge Analysis) toolkit [11]. In addition, a comparative analysis was performed to determine the algorithm that produced the best performance.

Further to the above, the research presented in Chapter 7 of this thesis also investigates time-series analysis of ground level ozone concentrations. In time series analysis both univariate and multivariate approaches are investigated with the use of both popular but simple machine learning algorithms (e.g., NN and SVM) and more advanced Ensemble Learning algorithms, such as Random Forests and Bagging. The analysis considered above is forecasting future values ozone concentrations based either only on the past ozone concentration values (this is univariate analysis) or together with also other

parameters that are known to have a direct impact on formation of ozone such as meteorological parameters and gases of known concentrations.

Two different experimental datasets (i.e. regression data) are used to support the investigations carried out in this thesis. The first dataset was obtained from Sohar University, Oman, which used a DOAS instrument (see Chapter 3) to gather the environmental data as well as concentrations of other gases known to cause ozone. The second, more comprehensive dataset was obtained from UK's biggest air quality dataset produced and made available by DEFRA. Details of the two datasets and their pre-processing are discussed in detail in Chapter 4.

Figure 1.1 illustrates a generic system for modeling the ground level ozone concentration. The system consists of three main stages namely, data collection, data preparation (data preprocessing) and modelling using machine learning algorithm. The data preprocessing stage consists of several steps which can be briefly listed as below:

1. Collect raw data
2. Remove missing values and outliers
3. Transform data if needed to suitable formats
4. Select features/attributes that best describes the data
5. Use data in creating the model

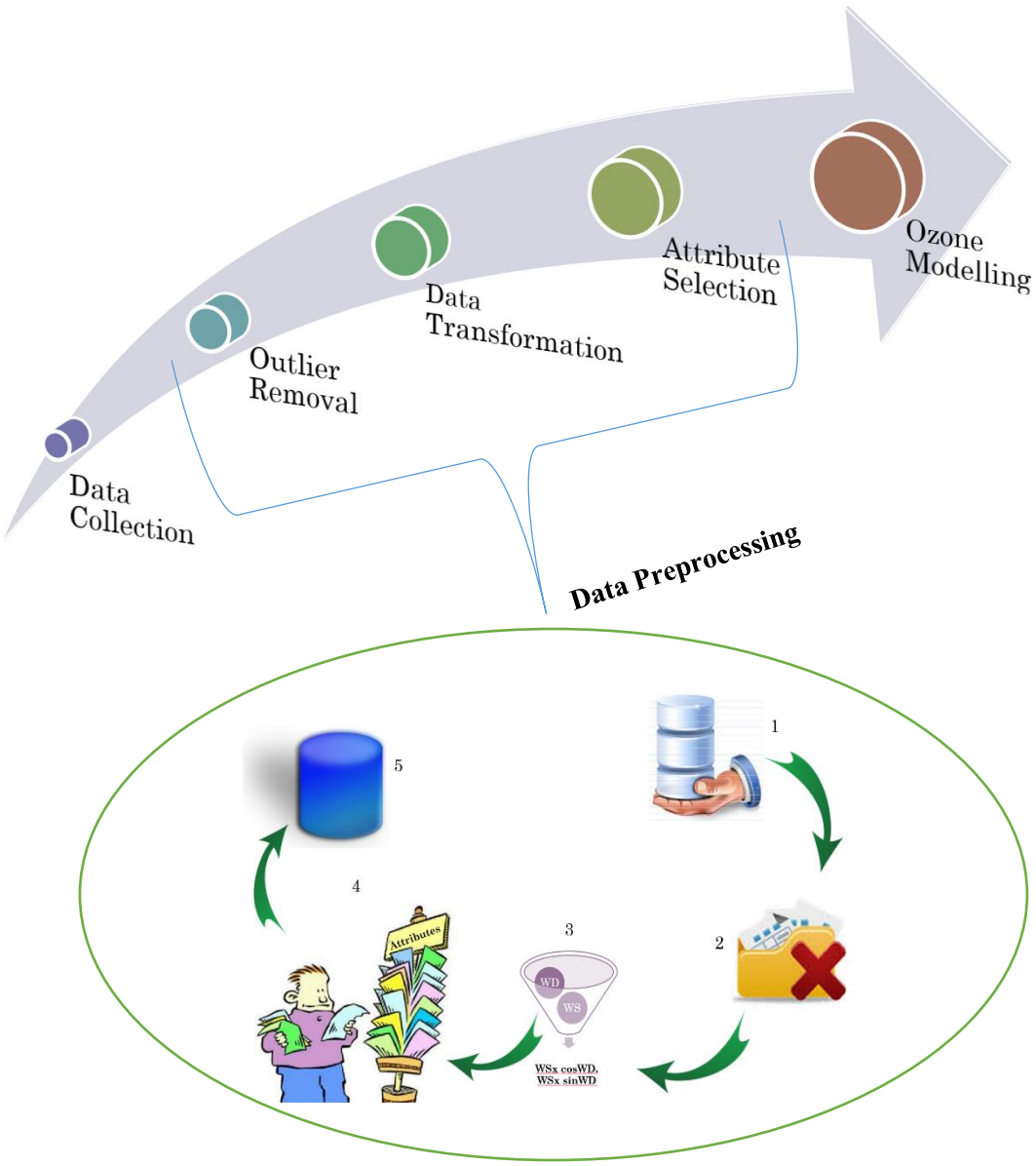

Figure 1.1: Generic System of Ozone Modelling

For the implementation of the machine learning algorithms used in the spatial and time-series prediction of ozone concentrations in this thesis, the popular data mining tool WEKA [12], is used. In addition to the implementation of a large number of popular and advanced machine learning algorithms, the software also implements data cleaning, feature selection, validation performance measurement tools. Throughout the research conducted and presented in this thesis such additional tools have been identified and used to support the rigourous investigations carried out within this thesis.

1.2. Research Motivation

The motivation for the research conducted within the context and scope of this thesis was derived from a comprehensive literature review conducted on atmospheric air quality prediction, modelling and monitoring that identified gaps in knowledge and associated research. The literature review conducted revealed that the existing research on air pollution measurement and analysis was largely limited to using traditional machine learning techniques (ANNs and SVMs) and statistical approaches. Only a few attempts have been made to utilise state-of-the-art machine learning approaches. Such studies have been limited in scope and rigour, and does not clearly conclude and recommend the best approaches to be used for a given dataset and type of analysis that is being carried out. Therefore, the research presented within the context of the work presented in this thesis studies and examines rigourously the ability of several ensemble learning methods to model, predict and forecast, ground level ozone concentrations.

Whilst rigorously comparing and contrasting the performance of a large number of machine learning algorithms in modelling, prediction and forecasting of ground level ozone concentration, the research conducted within this thesis gives a particular emphasis to the use of Ensemble Learning approaches. This particular aspect of investigation has been motivated by the research conducted by [13] who presented an empirical study about ensemble learning and explained why ensemble learning algorithms often perform better than any of the single classifiers in any general application area. Three fundamental reasons for constructing a good ensemble were discussed by Dietterich [13]:

(1) Statistical: ensemble algorithms can solve the problem of having a relatively large number of attributes compared to the number of instances in the dataset.

(2) Computational: in ensemble techniques the full input dataset is divided into smaller datasets allowing the search to avoid the so called *local minima problem*. In contrast a single base classifiers may risk getting stuck at a local minima due to the full dataset being considered as one.

(3) Representational: the mechanism of taking the sum of weighted functions in the ensemble methods can expand the space which it represents, while the single classifiers are limited by one representative function. However, it is noted that this can be provided by some other single classifiers, such as neural networks and decision trees, if they have been given enough training data.

Dietterich [13] also compared the three most common ensemble methods: Bagging, AdaBoost and Randomised Tree approaches. The results showed that adaBoost gives the best result in a noise free data set, while the performances of Bagging and Randomised Tree are very similar. However, randomisation has shown better result than Bagging in cases of very large datasets. For the purpose of deep analysis, the author added 20% artificial noise to his dataset. Hence, the performances of Bagging and adaBoost changed. The results showed the superiority of Bagging when compared to the under-performance of adaBoost and randomisation. This research and the associated observations have motivated the research conducted in this thesis which particularly focuses on Bagging, as an Ensemble Learning algorithm for the modelling, prediction and forecasting of ground level ozone concentrations.

1.3. Aim and Objectives

The research presented in this thesis aims to investigate the most effective and accurate modelling, prediction and forecasting approaches for the determination of ground level ozone concentration, using state-of-the-art machine learning algorithms, in particular, ensemble learning algorithms.

The following are the key objectives of the proposed work:

- To investigate the used of popular single base learning algorithms and ensemble learning algorithms in time series modelling and forecasting of ground level ozone concentrations.
- Examine the effective use of parameter based optimisation of learning algorithms to optimise the performance of time series modelling and forecasting approaches.

1.4. Contributions of Research

The research conducted within the remit of this thesis has resulted in the following original contributions to the area of atmospheric air quality modelling, prediction and forecasting:

1. Chapter 5 carries out a comprehensive investigation into the use of a significant number of machine learning algorithms for the modelling and prediction of ground level ozone concentrations based on meteorological parameters and concentrations of other gases known to cause ozone via degradation, due to natural phenomena. The performance of 16 different single base learning algorithms ware compared against a number of key ensemble learning algorithms including Bagging, Voting and Stacking used under default model parameter settings. The experiment's results have shown the ability of ensemble learning algorithms to significantly improve the prediction accuracy.

   The research outcomes of this work has been submitted as a journal paper to the Big Data Research Journal, Elsevier under a title "*The Use of Meta-Learning Ensemble Algorithms for the Prediction of Ground-Level Ozone*".
2. Chapter 6 carries out a detailed investigation into the performance optimisation of the machine learning algorithms recommended in Chapter 5 to perform best under their default parameter settings and non-optimised procedural operation. In particular parameter based fine tuning / optimisation of the popular single learner algorithms MLP (Multi-Layer Perceptron) and SMOreg (Support Vector Machine), single layer Ensemble Learning algorithm, Random Forest and multi-layer Ensemble Algorithms, Bagging has been investigated. The approaches that can be used to provide optimal performance of the advanced Ensemble Learning algorithms have been recommended. In addition, the use of feature/attribute/input, selection/reduction approaches to improve the performance accuracy of algorithms investigated, have been proposed and investigated in detail.

The original research outcomes and contributions of Chapter 6 has been published as a conference paper "*Al Abri, E.S., Edirisinghe, E.A. and Nawadha, A., 2015. Modelling ground-level ozone concentration using ensemble learning algorithms. Proceedings of the International Conference on Data Mining (DMIN), 27th-30th July 2015, Las Vegas, USA, pp.148-154*".

3. Chapter 7 carries out a detailed investigation into the use of popular single base learning algorithms and ensemble learning algorithms for time-series forecasting of ground level ozone concentrations. Both univariate and multivariate forecasting options have been considered. A number of conditions that can affect the forecasting performance of different machine learning algorithms have been investigated and recommendations are made for obtaining optimal performance from the investigated algorithms. The original contributions of this chapter will be submitted as a conference paper to *13th International Conference on Machine Learning and Data Mining MLDM 2017*.

1.5. Thesis Overview

For clarity of presentation, the thesis has been organised into eight chapters, as described below:

Chapter 1 provides an overview of the context of the research, identifies the research gaps and briefly presents the fundamental research focus of this thesis. The chapter also highlights the main contributions of the thesis and a chapter overview.

Chapter 2 reviews the existing literature on air quality monitoring and the work presented in literature that uses machine learning techniques for predictive modelling in areas outside environmental air pollution monitoring. The chapter concludes with a critical analysis of current literature, thus identifying research gaps in the subject area.

Chapter 3 focuses on providing background knowledge related to the research presented in this report. The chapter includes a brief description of the DOAS instrument that was used to gather one of the two datasets used in the experiments. The chapter also presents various machine learning algorithms that will be used for modelling Ozone

concentration in the forthcoming chapters and the evaluation criteria that will be used for the comparison of algorithmic performances. Furthermore, data pre-processing/cleaning filters used for data conditioning prior to modelling is also discussed.

Chapter 4 presents the two datasets used in the research conducted, namely the Sohar, Oman dataset and the DEFRA, UK dataset.

Chapter 5 investigates the use of a comprehensive set of single base learning algorithms and ensemble learning algorithms in the prediction of ground level Ozone concentrations, based on meteorological parameters and concentrations of other gasses known to impact the formation of ozone due to natural phenomena. The investigations are limited to identifying the best algorithms for further detailed investigation in Chapter 6.

Chapter 6 investigates the use of Ensemble Learning algorithm, Bagging and Random Forest in detail for modelling and prediction of ground level ozone concentration, making use of the popular single base learning algorithms such as Neural Networks and Support Vector Machines as the base layer algorithm. Parameter based optimisation of such algorithms and the use of feature/attribute reduction algorithms to further improve performance is also investigated in detail.

Chapter 7 investigated the use of machine learning algorithms in the univariate and multivariate forecasting of ozone concentrations in the atmosphere. The investigations have been limited to making use of the best performing single base learning algorithms and ensemble learning algorithms.

Finally, Chapter 8 concludes the research presented in this thesis, summarising the work carried out, making overall conclusions and identifying future directions for research and development.

CHAPTER 2

Literature Review

2.1. Introduction

Air quality monitoring is one of several crucial tasks carried out in the area of environmental science and engineering. Accordingly, the development of air quality predictive models can be very useful as such models can provide early warnings of pollution levels increasing to unsatisfactory levels. Such models can also be used to determine the causes and sources of pollution when combined with other data/information.

During recent decades, the harmful effects of air pollution on the environment and human health have been clearly noticed. Therefore, predictive models that can help monitor atmospheric pollution is required. In building such models, numerous attempts presented in literature have employed common data mining techniques as statistical tools. Due to the complexity and the non-linearity of air quality data, as first proved by [14], environmental researchers have popularly used Artificial Neural Network (ANN) and Support Vector Machine (SVM) algorithms for modelling atmospheric ozone concentration. These algorithms (ANN and SVM) have shown promising results compared with simple tree structured classification algorithms such as, CART, J48, and M5Rule. However, there exist more recent and advanced machine learning techniques that have not been used and investigated in the area of air quality modelling and prediction. These techniques are Bagging, Boosting, Staking, Voting and Random Forest, which are a family of ensemble learning algorithms. These ensemble learning algorithms have been experimented within areas that are beyond air quality management, where they have been shown to outperform the traditional, more popular algorithms, such as ANN and SVM approaches. On other hand, modelling air quality has been examined using ensemble techniques as statistical approaches in [15]–[18]. These researches had proved the ability of ensemble technique to outperform the single base techniques. However, the investigation of statistical approaches is beyond the

discussion research context of this thesis as machine learning based approaches have been proven to outperform statistical approaches, significantly. Therefore, in this research, an intensive investigation is carried out on the performance of a comprehensive set of data mining techniques (single based and ensemble learning) to build predictive models for atmospheric ozone concentration. Although existing literature has few examples of the recent algorithms being investigated for modelling atmospheric pollution, such research is limited to testing one or two of the approaches, without any comprehensive, in-depth analysis as to explaining their performance.

This chapter provides an in-depth literature review of existing work on the use of machine learning based approaches for atmospheric air pollution modelling and prediction in general (see Section 2.2) and in areas outside, such as in bioinformatics and marketing.

For clarity of presentation this chapter has been divided into several sections; each will introduce existing literature on using machine learning algorithms in different fields of study. Section 2.2 provides information about the data mining techniques that have been used to predict the ozone concentration level in the atmospheric pollution monitoring area, and is followed by Section 2.3, which presents the use of ensemble methods to build predictive models in other application areas. Finally, Section 2.4 summarises the chapter, making recommendations regarding the need of comprehensively investigating the use of ensemble classifiers in the modelling of atmospheric ozone concentrations.

2.2. Monitoring Atmospheric Air Quality

Almost all recent studies in atmospheric air quality prediction and modelling have employed an ANN to build a predictive model. The reason goes back to the proven properties of typical atmospheric ozone concentration datasets. These datasets have been described by [19] as very complex and show non-linear relations between the factors which effect the production of ozone. Abdul-Wahab and Al-Alawi [20] proved that an Artificial Neural Network (ANN) model can handle this complexity. Therefore, most researchers in this area have limited their studies to the use of ANN and SVM. The

literature below presents several studies in which ANN and SVM have been used. It is noted here that only a few attempted to employ different machine learning algorithms and carried out comparisons with ANNs. In addition the sensitivity analysis of the model was illustrated in some of the research along with using machine learning within air quality forecasting.

2.2.1. Literature Review

A Rigorous Inter-Comparison of Ground-Level Ozone Predictions[21]**.**

The paper presented a comprehensive study by comparing the performance of 15 statistical techniques for ozone concentration forecasting in Europe. These techniques were examined using 10 different datasets containing different meteorological and emission conditions. A Neural Network was amongst these techniques, and was studied and compared with the rest of the statistical tools. However, the main focus was on statistical tools only. The paper concluded that a satisfactory result was obtained only by the methods which could handle non-linear data. Thus the Neural Network was proven to perform best. The research demonstrated the ability of Neural Networks to work on non-linear datasets. Consequently, the authors of most of the papers in this section cited this study as the reason for using Neural Networks in their research.

Systematic Approach for the Prediction of Ground-Level Air Pollution (around an Industrial Port) Using an Artificial Neural Network [22]**.**

The study conducted in this thesis was carried out in the City of Sohar, Oman, where part of the data for the proposed research was gathered. The study proposed to develop models for daily predictions of CO, PM10, NO, NO_2, NOx, SO_2, H_2S, and O_3. The training of the prediction models was based on the Multilayer Perceptron method with the Back-Propagation algorithm, and showed very high concurrence between the actual and predicted concentrations. In addition, the research investigated the MLP model's sensitivity to variation of epochs cycle (trial and error technique adopted to try different adjustments).

Forecast of Air Quality Based on Ozone by Decision Trees and Neural Networks [23]**.**

In this research the Authors concluded that the Multilayer Perceptron Neural Network and algorithms such as Random Forest are capable of predicting O_3 with a similar accuracy. These algorithms will be used in this study with additional attributes such as BTX and solar radiation. Further the use of Support Vector Machines and Bayesian Networks were also considered.

Prediction of Missing Data for Ozone Concentrations Using Support Vector Machines and Radial Basis Neural Networks [24]

The paper studied ozone concentration data in two seasons (summer and winter) to forecast the ozone level. The work proposed used Support Vector Machines (SVM) and Artificial Neural Networks (ANN) to predict the ozone level in two phases: in an hourly basis and one on a weekly basis. The data which was used in the experiment was obtained from the Republic of Macedonia, during the year of 2005. Clean data was selected from the dataset for 10 days in a row for two months (August and December). In addition, the experiment focused only on four input parameters, nitrogen dioxide (NO_2), ozone (O_3), temperature and humidity. They used the WEKA package to build three different models. The first two models used a Support Vector Machine with two different kernels (Polynomial and Gaussian) and the third model used a Artificial Neural Network (ANN) with Radial Basis Function (RBF). The experiments demonstrated that the results provided by SVM with both kernels were better than when using ANN. In conclusion, the author recommended a further study with more meteorological variables.

Neural Networks for Analysing the Relevance of Input Variables in the Prediction of Tropospheric Ozone Concentration [25]**.**

This study employed a Neural Network to build a prediction model for ozone concentration levels and obtained the relationships between the relevant variables.

This experiment included more input variables than the above study: two vehicle emissions (NO and NO_2) and five meteorological variables (temperature, wind speed, relative humidity, solar irradiation and pressure). Furthermore, a sensitive analysis to determine the relevance of the input variables was carried out in this work. The results showed the complexity, non-linearity and time dependency of the ozone concentration prediction mechanism.

Prediction of Ozone Concentration in Tropospheric Levels Using Artificial Neural Networks and Support Vector Machines at Rio de Janeiro, Brazil [26]**.**

The paper aimed to analyses the behaviour of the input variables of the air quality monitoring dataset. In addition, the implantation of non-linear regression methods, SVM and ANN, were used to build a predictive model for ozone concentration. Both techniques used three different datasets created for the purpose of this study. These databases varied based on season, pollution source, and weather conditions. However, the first two datasets were taken from different places, while the third was produced by merging the first two sets. Hence, an indication of a dependent relationship between ozone concentration and other pollutants and meteorological conditions was obvious in the results. Moreover, the study examined and showed the non-linear relationship between ozone and other input factors. In addition, the resulting predictive model using SVM and ANN was found to be very consistent with actual observations.

Assessment and Prediction of Tropospheric Ozone Concentration Levels Using Artificial Neural Networks [20]**.**

The research employed a ANN to predict the ozone concentration level in an urban area using pollution and meteorological measurements. The relationship between ozone and other ambient air measurements was studied as well. In the work proposed the authors focused on a summer dataset, as the ozone concentration is very high during that period. Therefore, three models were developed for different investigations, with different input variables. The first two models investigated the

factors which control the ozone concentration. However, the first model focuses on a 24 hour period, while the second model looked at the daylight period, when the highest ozone concentration was recorded. Moreover, the third model predicted the daily maximum ozone concentration level. The experiment confirmed the conclusion of other researchers, i.e. that the ozone concentration is high during the summer season. In addition, the results showed the dependent relation between ozone and other input variables (air pollution and meteorological variables). The study also shows the ability of a ANN method to predict and model the ozone concentration.

Hourly Ozone Prediction for a 24-H Horizon Using Neural Networks [27]**.**

The paper aimed to verify the presence of non-linear dynamics in the ozone concentration time series. They presented a study of hourly ozone prediction for 24 hours per day of a whole year. Therefore, two ANN structures (dynamic and static models) were adopted to build the model. The dynamic model of ANN is represented by a cascade 24 Multilayer Perceptron (each with its own output). In other words, the output of the previous flow will be fed as an input of the next one. On the other hand, the static approach has only one layer with 24 outputs. The authors have used the ANN to implement the experiment based on previous studies which have used and proven the ability of ANN to predict the Ozone concentrations. For dynamic model the model used the previous forecasting ozone along with past 24 hour (24 lagged variables) of meteorological parameter and NO_2, while the static used only the past 24h of meteorological parameter and NO_2. The result showed both model have a comparable result which indicate there is no dynamic nonlinear relation on ozone time series. In addition, the author employed sensitivity analysis to test the generalization ability of the two architectures (static and dynamic model). After the inclusion of an optimisation procedure to the two models, the research has introduced the perturbation to input values of the test set from 10 to 40% while the training set was kept as the original. Moreover, another examination has been applied when the same perturbation amount was introduced to training the set while fixing the test set. The results of the test have shown that small changes in the input impact the output.

Learning Machines: Rationale and Application in Ground-Level Ozone Prediction [28]**.**

The paper aimed to produce a study which could be considered a benchmark for further research into atmospheric air quality prediction. They presented the use of MLP with different methods to predict the ozone level. MLP suffers from two main problems, i.e. over-fitting and local minima; a number of researchers have subsequently tried to overcome these problems. Lu et al. [29] applied Practical Swarm Optimisation (PSO), while [30] tried MLP with automatic relevance determination. However, no studies have successfully addressed the two problems simultaneously. Therefore,[28] proposed the use of SVM to address the MLP problems mentioned above. In this study, SVM performance was examined and compared against MLP. The result illustrated that the MLP is better in the sense of risk immunisation, while the SVM was better that the MLP on structural issues.

Multiple Linear Regression And Artificial Neural Networks Based on Principal Components to Predict Ozone Concentrations [31]**.**

The research conducted a study to build a model to predict the hourly ozone concentration. The study employed the feedforward ANN and multiple linear regression with and without Principle Components (PC) as input. Principle Component Analysis (PCA) was used to transfer and reduce the number of predictive variables to new input variables, i.e. a set of Principle Components. In addition, the study investigated the influence of each environmental factor on ozone concentration using statistical analysis. The authors compared the model designed with PC against the original variables. The result showed the ability of Principle Components to improve the prediction of the model. Besides this, PC can reduce the data complexity and co-linearity of the dataset.

Forecast Urban Air Pollution in Mexico City by Using Support Vector Machines: A Kernel Performance Approach [32]**.**

The paper used SVMs to model the concentrations of air pollutants in the city of Mexico. The study utilised several kernels (Gaussian, polynomial and Spline) and compared their performances against each other. The SVM model provided good accuracy in modelling the concentration of the gases ozone, nitrogen dioxide and PM10. They concluded that SVM with a Gaussian kernel provides better performance than the others. However, when using SVM, the Gaussian kernel may not be satisfactory in reality due to its high computational cost.

Development and Evaluation of Data Mining Models for Air Quality Prediction in Athens, Greece [33]**.**

The authors examined the performance of multiple data mining algorithms for modelling air quality. The experiment was implemented using the best known software package for machine learning (WEKA). 84 different models were built from almost all the algorithms implemented in the software. Different statistical tests and indexes were used to evaluate the models. The best performances in classification were obtained from J48, LMT, OneR, Decision Table and REPTree. These algorithms fall under the Tree and Rule categories. In addition, the best algorithms under a regression model are M5P, REPTree, and M5Rule. Meanwhile, the worst results were obtained from the models based on SMOreg and linear regression. However, this work did not consider any of the ensemble classification algorithms which were not implemented within WEKA as the time the study was conducted.

Identifying Pollution Sources and Predicting Urban Air Quality Using Ensemble Learning Methods [9]**.**

The paper conducted one of the few atmospheric environment studies that considered the use of ensemble learning (Bagging and Boosting) for constructing an air quality model. These models were utilised to differentiate air quality during the

different seasons and predict an air quality index. In addition, the study used PCA to identify the main and relevant air pollution sources. Hence, the experiments illustrated a noticeable enhancement in the accuracy of ensemble learning, compared to SVM. Moreover, the PCA identified fuel combustion and vehicular emission as major air pollution sources in the city where the study was carried out.

Forecasting Summertime Surface-Level Ozone Concentrations in the Lower Fraser Valley of British Columbia: An Ensemble Neural Network Approach [10]

Another study that attempted to compare the ensemble methods and the single base learners was presented in this paper. The study built a predictive model to find the maximum average ozone concentration in daylight hours during the summer season. The developed models initially used Multilayer Perceptron (MLP) and Multiple Linear Regression (MLR). To support the experiment, data was taken from 10 different stations for a period of five years. However, both methods suffered from instability and over-fitting problem. Therefore, use of Bagging (one of the ensemble learning techniques) was proposed to improve the stability and accuracy of the designed model. Later, a comparison between bagged MLP (Bagging with MLP as base classifier) and bagged MLR was carried out and compared against the individual models of MLP and MLR. The result showed the ability of Bagging to improve the stability and accuracy of MLP. On the other hand, the result of Bagging was disappointing when using the MLR model. However, bagged MLP outperformed bagged MLR in all the stations.

Assessment of Adding Value of Traffic Information and other Attributes as Part of its Classifiers in a Data Mining Tool Set for Predicting Surface Ozone Levels [8]

This paper utilizes the Sohar dataset which has been employed in this research (see Chapter 4), however, the aims was to examine the effect of traffic information on ozone predication. Several prediction models were contrasted and compared to model the ozone concentration. These models included the single base algorithms

(M5P, REPTree, Kstar, M5Rule, IBK, SMOreg, MLP, Decision Table, LWL, Decision Stump, and RBFNetwork) and ensemble learning algorithms (Bagging, Random Subspace, Regression by Discretization, and Addictive Regression). The results indicated that Bagging was the best classifier based on the values of evaluation indexes (CC, MAE, RMSE, RAE, and RRSE). In addition, the paper extended the research to optimise the parameters of two models (Bagging and M5P). However, the results shown that no significant improvement of the accuracy is obtainable. More experiments on Bagging employed during the day time and using traffic information and compared it without this information. The results showed improvement in the productive model of ozone.

Effective 1-Day Ahead Prediction of Hourly Surface Ozone Concentrations in Eastern Spain Using Linear Models and Neural Networks [34]

The paper used ANNs in forecasting the ozone concentration, 1 day ahead. The authors used data over a 4 years period. The paper aims to use the past and previously predicted values of the inputs to forecast the ozone concentration for one day a head. The authors used the autoregressive moving average with exogenous input (ARMAX), MLP, and finite impulse response (FIR) Neural Network. The paper used an approach for forecasting the hourly ozone that differs from the previous study (where they used the future information of the covariates in order to predict future of ozone) in two aspects (1) they have used the past information of meteorological parameters and ozone as inputs to the model, and not using the currant values of the variables. (2) used the previous forecasted ozone concentrations to predict the current ozone concentration. The authors employed the lag information as an input to the model.

Urban Air Pollution Monitoring System with Forecasting Models [35]

Another recent paper has tested three different classification algorithms, SVM, M5P and ANN, for forecasting polluted gases O_3, NO_2 and SO_2. The paper has built

two types of models for each gas: (1) Univariate model where the model was built using the target gas only, (2) Multivariate model where other features of gases and meteorological parameters are employed. The paper has used different lag variables (8 and 24) to forecast 1, 8, 12, and 24 hours ahead. The results have indicated the M5P outperform SVM and ANN for all the gases in all the forecasting steps they have tested. The paper did not study the use of ensemble learning algorithms for forecasting ozone concentrations.

Neural Network Modelling and Prediction of Hourly NO_x And NO_2 Concentrations in Urban Air in London [36]

This paper aimed at building forecasting models for hourly observation of NOx and NO_2 concentrations using MLP. The model was compared with a linear regression based model (LR). The results have indicated the power of the MLP to deal with complex patterns. Different MLP models and LR models were constructed using different inputs. Some of these models have introduced the historical information of (NO_X/NO_2) with two different lagged variables (lag-1 or lag-24). The results indicted the use of lag-1 is not practical for forecasting, while the lag-24 provided some reasonable predicted episodes. Furthermore, the tests have shown that using lagged variables, improves the model performance.

Hybrid Model for Urban Air Pollution Forecasting: A Stochastic Spatio-Temporal Approach [37]

The paper employed a ANN to concentrate the PM10 forecasting model using the lagged variables and meteorological parameters. The input parameters to the model were the previous day mean of NO, NO_2, CO and PM10 concentration, mean of meteorological parameters for previous day, and PM10 concentration of the current day. A correlation analysis (backward stepwise regression) adopted to determine the parameters that has the most significant impact to the PM10 forecasting model. These analyses enabled the optimisation of the ANN model.

Multilayer Perceptron and Regression Modelling to Forecast Hourly Nitrogen Dioxide Concentrations [38]

The paper presented forecasting models for NO_2 using MLP and multiple regression (MLR) models that predicted 24 hours ahead. The authors have used hourly data of NO_2 concentration, meteorological parameters and a linear presentation of the daily and weekly periodic (sine and cosine of each information) values. Both selected learning models were optimised. Several models were constructed using different input parameters employed by both learning algorithms. The Seasonality (t+24), sine and cosine of daily and weekly cycles, were used as an inputs for some of the models. The result showed that a good forecasting model can be achieved using MLP. Furthermore, the best MLP model was obtained when the input parameter used meteorological values, traffic information, seasonality t+24, NO_t and $(NO_2)_t$ concentrations, while the best MLR model was obtained when using meteorological values, traffic information, and $(NO_2)_t$ concentrations.

Forecasting Seasonal Time Series with Neural Networks: A Sensitivity Analysis of Architecture Parameters [39]

The paper aimed to investigate use of MLP in building a forecasting model within different seasonal and trend components. The paper discussed the best MLP architecture to build the forecasting model (OPTIMAL Model) and examined the sensitivity of different MLP to forecast seasonal dataset. The MLP topology was tested in respect to input node, number of layers and activation function. The authors have focused on altering the investigated parameters, while keeping the other parameters set to its generic architecture. The result has shown that the variation of the input and the MLP parameters were impacting on the model's performance.

Linear and Nonlinear Modeling Approaches for Urban Air Quality Prediction [40]

The paper aimed at predicting the air quality and examined the sensitivity of the input variables by observing the effectiveness of eliminating some of the inputs. Three different types of the ANN models were investigated, namely, the multilayer perceptron network (MLPN), radial-basis function network (RBFN) and the generalized regression neural network (GRNN). These models were used to construct models for RSPM, NO_2, and SO_2, a predictive model for air quality using meteorological and urban air quality datasets. Hence, for sensitivity analysis the study tested and compared all possible combinations of the input variables that contain groups of one, two, three, four, and five variables. The performance of these tests analysed with an optimal GRNN structure (since the GRNN models outperformed MLPN and RBFN models). The analysis of sensitivity of the models (RSMP, SO_2, and NO_2) indicated which of the input groups, were performing the best in each model.

Establishing Multiple Regression Models for Ozone Sensitivity Analysis to Temperature Variation in Taiwan[41]

The paper studied the sensitivity of the concentration of ground level ozone to temperature changes in Taiwan. Multiple regression models were built using an hourly dataset including ozone, PM10, temperature, relative humidity, wind speed, wind direction, and rainfall. Data were analysed from 2000-2009, where the data from 2000-2009 were used for descriptive statistics and data 2004-2009 were utilised to construct the model. A dataset of 2070 records for surface ozone concentration has been tested for its sensitivity to temperature by replacing the original value with different values of temperatures ±1 C, ±2 and ±5. The indictors of this test (ozone sensitivity to temperature) were the 75th percentile (yearly maximum) and number of ozone exceedance. In addition, a sensitivity analysis was also carried to the 2030 predicated data. The paper concluded there are a positive correlation between ozone sensitivity and temperature variation. Hence the paper focused on future

temperatures and the analyses applied to 2030 predicated data results has indicated that ozone sensitively was strongly associated with ozone seasons. A result which agrees with the findings of other existing studies.

Exploring the Utility of the Random Forest Method for Forecasting Ozone Pollution in SYDNEY [42]

This paper examined the performance of the random forest in forecasting the ground level ozone pollution. The performance of Random Forecast was compared with single tree (CART) classifier. The paper has developed two classification models each using decision tree (CART). In addition, the Random Forest has implemented to go through two phase of modelling to get better accuracy. In the first phase a classic random forest classifier was applied to the training set and to build the forecasting model. Phase two employed the random forest classifier but using a new training set from the phase one's forecasting model. This training set consists of all instances which are incorrectly classified. This idea was adopted due to the use of the boosting strategy to boost the result. The paper concluded that results of using Random Forest outperformed the single decision tree in general.

2.2.2. Research Gap

For the last 10 years, atmospheric researchers have used ANNs or SVMs as statistical tools to assist them in monitoring air quality. The preference for these tools is based on the experiments and conclusions of previous research in this area. Meanwhile, there are more advanced techniques in machine learning (ensemble learning) which have demonstrated their ability for more efficient prediction in other areas. Unfortunately, only a few researchers have attempted to use these techniques, namely [8], [9], and [10] (discussed in the above section). A comprehensive investigation into the use of ensemble learning methods with different combinations of base classifiers has not been found in literature. This will be a key focus of the research conducted and presented within this thesis.

Section 2.3 presents existing studies in other application areas (i.e. outside machine learning), which have successfully employed the ensemble learning algorithms, obtaining better results in prediction ability.

2.3. Use of Ensemble Learning Algorithms in Other Areas

Section 2.2 provided an insight into studies of air quality modelling focused mainly on modelling, prediction and forecasting of ozone concentration. These studies used available statistical and machine learning tools to build the desired models. Looking at the nature of data related to air quality, atmospheric researchers have used ANN and SVM algorithms in their design and implementation of models. The reason for this comes from the fact that both of the algorithms are known for their ability to model non-linear datasets, e.g. atmospheric and/or ground level ozone concentration. On the other hand, machine learning now has more advanced techniques (e.g., ensemble learning technique) which have the ability to work with massive and complex datasets. However, ensemble learning algorithms have not been employed widely and rigorously in the modelling and prediction of ozone concentrations. Therefore, this section reviews the use of ensemble methods in areas outside that of atmospheric pollution monitoring.

2.3.1. Literature Review

A Review of Ensemble Methods in Bioinformatics[43]

Yang et al. [43] presented a review of ensemble methods in the area of bioinformatics. This paper discussed the application of the ensemble methods in three topics in bioinformatics: (1) Classification of gene expression microarray data and MS-based proteomics data; (2) Gene-gene interaction identification, and (3) The prediction of regulatory elements of DNA and protein sequences. In all three applications, the authors demonstrated that the ensemble methods are more accurate than the single base methods. The study concluded that ensemble learning in general can achieve higher accuracy and stability than single base

algorithms.

Ensemble machine learning on gene expression data for cancer classification. [44]

Tan and Gilbert [44] studied the use of ensemble learning (Bagging and Boosting) to classify a cancer from gene expression data. The performance of ensemble learning was compared with a single algorithm, C4.5 Decision Tree. The improvement obtainable from ensemble learning algorithms was obvious. The study concluded that the ensemble learning techniques often perform better than single algorithms in classification tasks.

A study in the area of medical data analysis by [45] compared the performance of Bagging with 12 other learning algorithms. These algorithms are the most commonly used in real-world applications: Support Vector Machine, Neural Network learner–MLP, Naïve Bayes learner (NB), K-nearest-neighbors (KNN), PART, DecisionTable, OneR, C4.5 DecisionTree, J48, DecisionStump, RandomTree, REPTree, and Naïve-Bayes-Tree. The study used eight imbalanced medical datasets and different statistical tests for the evaluation of models. The result of comparing Bagging with a single algorithm did not show a significantly better result than some of the single learners, such NB, SVM, KNN, DecisionTable and DecisionStump. On the other hand, Bagging statistically performed much better than the remaining single learning algorithms, such as J48, RandomTree, OneR, MLP, and PART.

Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks [46]

The paper presented another comparison study. The study compared adaBoost (an ensemble method) and ANN in accurately predicting the corporate failure of European firms. The results showed that the ensemble method adaBoost outperformed the accuracy of ANN. In addition, adaBoost reduced the generalisation error by 30%, compared to the error produced by ANN.

Prediction of Oil Prices Using Bagging and Random Subspace [47]

Gabralla and Abraham [47] presented a comparative study between several single based learning and ensemble learning (Bagging and Random Subspace) algorithms. The paper aimed to find the most accurate model to predict oil prices. Bagging and Random Subspace were examined using six different base classifiers, namely: Multilayer Perceptron (MLP), Isotonic Regression, Sequential Minimal Optimisation for Regression (SMOreg), Multilayer Perceptron Regressor (MLP Regressor), Extra Tree and Reduce Error Pruning Tree (REPTree). Each of these six models was utilised individually to construct a prediction model. Hence, the results were compared with the corresponding Bagging and Random Subspace algorithms. The experiments showed that the ensemble methods enhanced prediction accuracy except for MLP, with Random Subspace achieving the best results of the adapted techniques.

An ensemble method for predicting biochemical oxygen demand in river water using data mining techniques. [48]

A study proposed by [48] employed Bagging to predict the biochemical oxygen demands in river water. The authors proposed a model that used Bagging with K-star and compared it with seven other models. These models were ANN, Bagging with ANN, SMOreg, Bagging with SMOreg, Multivariate Regression, Regression by Discretization, and K-star. The experiments showed the improvement obtainable in accuracy when Bagging was employed.

Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. [49]

In the field of Electrical Power and Energy Systems, the paper presented a comprehensive study using most of the machine learning methods. The paper aimed to build a prediction model for a thermodynamic system. 15 predictive models were built and compared using several classifiers, namely Simple Linear Regression, Linear Regression, Least Median Square, Multilayer Perceptron,

Radial Based Function Neural Network, Pace Regression, Support Vector Poly Kernel Regression (SMOreg), IBK, Kstar, LWL, Additive Regression, Bagging with REPTree, M5P, and finally REPTree. As shown by the results of the experiments, Bagging provided the highest prediction accuracy.

The above works indicate that several attempts have already been made in areas beyond air quality prediction in the use of ensemble classifiers. This work has shown that ensemble classifiers outperform the corresponding single classifiers and that the ultimate answer to the question, which classifier works best, depends on the dataset. It is clear that different datasets, in particular from different application domains, are statistically different and this has a high impact on the variability of results.

2.4. Summary & Conclusion

As discussed in this chapter, several studies in the field of environmental science and engineering have focused their interest on constructing a model to deal with air pollution problems. Based on literature above, it can be seen that the methods used by environmental researchers to solve the problem have been rather limited. The majority of environmental researchers tend to use Artificial Neural Networks (ANN) and Support Vector machines (SVM) to model and predict ozone concentration [25], [28], [31], [32], [50]. However, some other researchers have attempted to apply ensemble learning (bagged ANN) and compared the performance with that of the single base learning algorithms, such as the work of [9] and[10]. Both [9] and [10] have proven that the use of Bagging shows improvement over the performance of single models. In the area of forecasting ozone concentrations into the future, the focus has been once again mainly the use of single learning algorithms.

On the other hand, several researches in other research areas such as bioinformatics, medicine and marketing, have successfully used ensemble learning methods to build predictive models [51]–[53] Despite the above conclusion, the use of ensemble learning for predicting ozone concentration was limited to few attempt of using ANNs. A comprehensive and rigorous study of using various ensemble learning algorithms, such

as, Random Forests, Bagging, Voting, Stacking etc. does not exists either within or outside the area of air pollution monitoring.

From the literature review above, a lack of research into utilising ensemble learning to predict ozone concentration was identified. In addition, an absence of a complete comparison between all machine learning methods for monitoring air quality was noticed. Therefore, in this work a comprehensive study has been conducted to examine the performance of ensemble methods against other 13 single based algorithms. The study will be conducted on an ozone concentration dataset using two different datasets and a predictive/forecasting model will be built. Moreover, a comprehensive analysis of the use of the Bagging classifier versus the other ensemble classifiers and all single classifiers will be carried out. Therefore, the proposed work can be used as a key reference for modelling air quality monitoring in the future.

CHAPTER 3

Research Background

3.1. Introduction

This chapter presents the background of research, on which the work presented in this thesis is built on. Theoretical explanations of the algorithms behind the machine learning techniques used, approaches used for data collection and techniques/metrics used for data analysis and evaluation will be defined and explained in detail as appropriate. In addition, the use of different filters and packages employed within the WEKA toolkit [12] are presented.

For clarity of presentation this chapter is divided into several sections. Section 3.2 presents the DOAS (OPSIS) instrument that was used by Sohar University, Oman, to gather the air quality data used in the experiments of Chapter 5 and Chapter 6. Section 3.3 demonstrate the different between machine learning and statistical models. Section 3.4 presents general information about the use of machine learning in data modelling and specific information about the particular learning algorithms used in modelling ground level ozone concentrations, in Chapters 5, 6 and 7. Section 3.5 presents the ten-fold cross-validation approach used in all experiments to validate the performance of the proposed models and Section 3.6 presents the validation metrics used. Section 3.7 presents the feature selection filters used and the optimisation algorithms used in adjusting the parameters of the training algorithms for optimal performance. Finally, Section 3.8 summarises the research background covered in this chapter.

3.2. DOAS (OPSIS) Instrument

A Differential Optical Absorption System (DOAS) [54] is used to measure the concentration of several gaseous species in the troposphere simultaneously. The system

records and evaluates the characteristic differential absorption of UV/Visible light source transfer, over a path of several kilometres. The DOAS technique is based on Beer-Lambert's absorption law, which is explained as follows: "*The quantity of light absorbed by a substance dissolved in a fully transmitting solvent is directly proportional to the concentration of the substance and the path length of the light through the solution*"[55]. Since the absorption spectrum property of each gas is unique, the concentration of each gas can be identified separately at the same time.

The DOAS technique is based on two main parts. The first part is responsible for sending a beam of light from a special source through a particular path. The light which is sent contains light of the visible spectrum, ultraviolet and infrared wavelengths. The second part is a receiver, which receives the light sent by the first part (the emitter). The receiver will transfer the light through an optical fibre to an analyser, as illustrated in Figure 3.1.

The analyser includes a computer, associated control circuits and a high-quality spectrometer. The spectrometer will use the optical grating to split the light into narrow wavelength bands. Subsequently, the light is transformed into electronic signals so that the computer can evaluate and analyse the light losses due to molecular absorption along the path. After several computer calculations, the instrument will produce a monitoring database, which includes the gaseous concentrations, with a high level of accuracy [54].

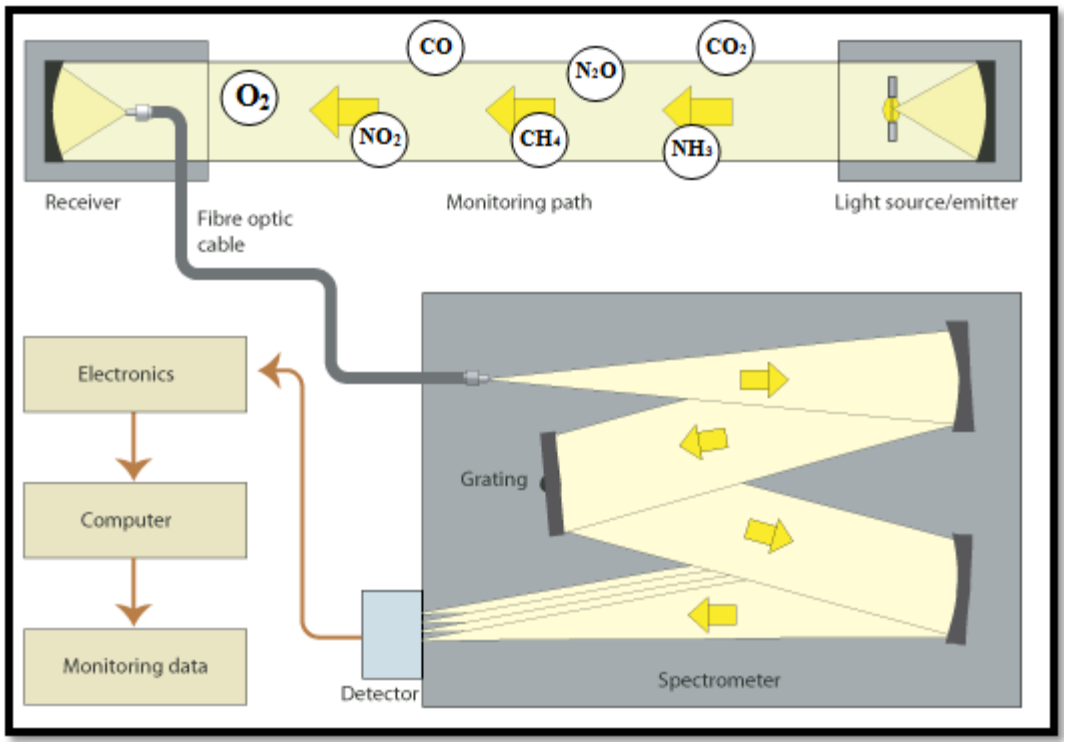

Figure 3.1: UV DOAS Technique [54]

The DOAS used by Sohar University, Oman that provided data for the experiments conducted in thesis in Chapters, 5 and 6, measures the concentrations of eight gases, namely SO_2, NO_2, O_3, benzene and (o-,m-,p)-xylene (refer to Table 3.1) .

Table 3.1: Sample of the dataset provided by OPSIS (Sohar University)

Record	DateTime	Sulphur Dioxide (SO2)	Nitrogen Dioxide (NO2)	Ozone (O3)	Benzene (C6H6)	Toluene (C7H8)	p-Xylene (C8H10(p))	m-Xylene (C8H10(m))	o-Xylene (C8H10(o))
1	04-01-13	-999	-999	-999	-999	-999	-999	-999	-999
2	04-01-13 1:00	3.589	17.902	39.901	6.162	12.551	6.02	5.677	42.097
3	04-01-13 2:00	2.549	12.858	36.98	10.049	16.559	5.151	6.862	35.563
4	04-01-13 3:00	2.707	11.785	34.844	6.233	19.229	7.59	6.758	32.874
5	04-01-13 4:00	2.681	7.035	44.41	7.094	11.843	5.938	5.608	33.723
6	04-01-13 5:00	2.738	24.482	34.59	6.917	8.104	5.309	6.709	33.064
7	04-01-13 6:00	4.203	68.765	14.12	8.259	26.228	4.851	6.847	25.519
8	04-01-13 7:00	6.918	65.055	19.863	11.462	47.001	4.799	6.54	30.905
9	04-01-13 8:00	4.594	33.758	28.78	17.398	27.151	5.396	5.444	34.869
10	04-01-13 9:00	8.294	37.592	44.105	9.841	19.69	5.833	5.654	47.725
11	04-01-13 10:00	10.013	32.001	65.23	7.642	5.174	5.766	6.156	47.475
12	04-01-13 11:00	11.286	21.158	78.667	6.826	9.777	4.961	6.889	50.703
13	04-01-13 12:00	7.459	12.573	88.073	16.37	16.872	8.916	8.34	36.892
14	04-01-13 13:00	7.22	10.391	96.685	8.236	9.005	5.361	6.167	39.215
15	04-01-13 14:00	6.475	8.742	106.078	9.696	15.038	4.749	7.323	34.121
16	04-01-13 15:00	6.823	11.757	106.602	11.342	3.499	6.618	7.707	40.261
17	04-01-13 16:00	7.416	22.77	95.71	17.409	7.856	8.421	7.09	37.939
18	04-01-13 17:00	7.623	26.095	79.386	14.747	18.475	7.888	6.684	39.471
19	04-01-13 18:00	7.56	58.699	60.833	8.088	8.458	5.331	7.347	44.574
20	04-01-13 19:00	7.354	82.522	40.549	11.888	35.377	6.314	6.078	55.082
21	04-01-13 20:00	6.975	59.254	42.866	11.817	87.232	5.047	6.056	26.517
22	04-01-13 21:00	5.936	83.087	30.081	11.11	34.852	5.23	6.512	44.139
23	04-01-13 22:00	4.967	44.13	46.299	6.035	13.798	4.768	7.328	34.757
24	04-01-13 23:00	3.177	19.788	59.15	5.765	6.752	5.043	6.178	35.656

3.3. Machine Learning V.S Statistic Analysis

A statistical model aims to develop a model that explains the data, while, machine learning develops a method to solve a problem. According to Witten et al. (2011, p.28-29) [11] there is no significant difference of principle between the two models. The concern of statistical is more toward the hypothesis testing. In contrast machine learning is more about developing the process to search for possible hypotheses

3.4. Machine Learning Techniques

Machine learning (ML) is the process of learning useful information from a large set of data, i.e., “big data”. This learning process leads to developing the capability to make intelligent decisions or predicting upcoming/future data. Therefore, ML has the ability to develop methods or tools that can be used to discover unseen patterns from given (i.e. seen) data to solve a particular task or problem. Subsequently, the built model can be

used to predict new data or information. Figure 3.2 illustrates the general process of machine learning.

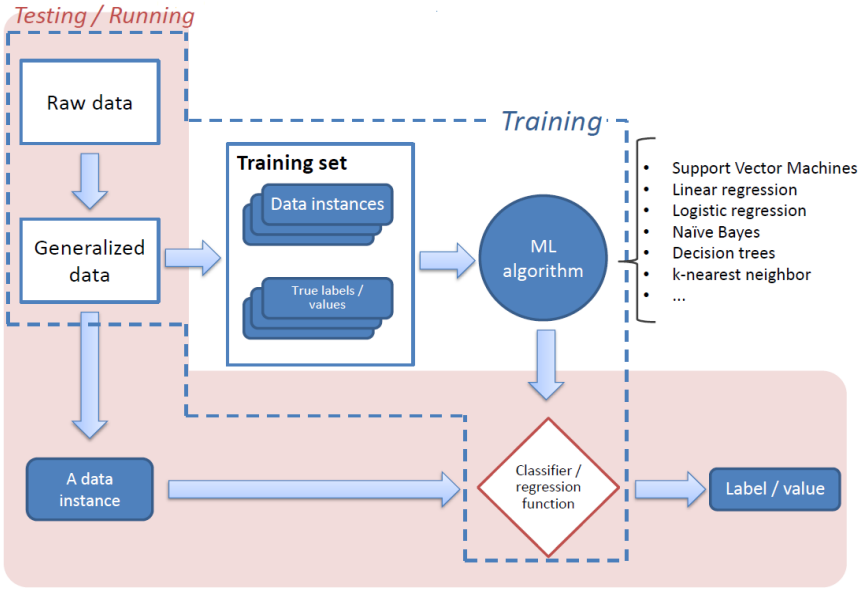

Figure 3.2: The Machine Learning Process (Supervised Learning)

It is worth mentioning that data mining tasks can be categorised into two types: (1) classification tasks, where the target class to be predicted is nominal and (2) regression tasks, where the target class is numeric (the work of this thesis). Not all the learning algorithms can handle the two categories [11].

A large number of machine learning algorithms have been proposed in literature. In general, these techniques can be divided into two categories: single learner algorithms and ensemble learner algorithms or meta learner algorithms. Figure 3.3 illustrates the main difference between the two methods. The details of each type are discussed in the subsections below.

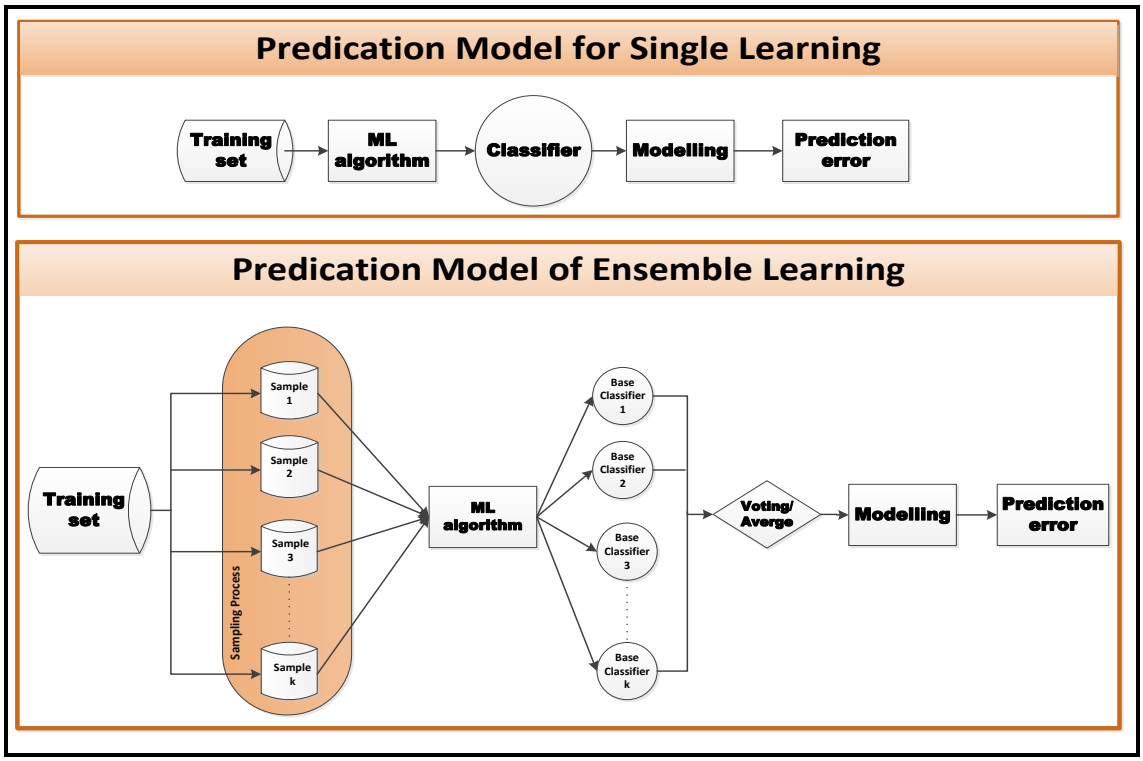

Figure 3.3:The processes of machine learning: Single Base Learner and Ensemble Learner algorithms

3.4.1. Single Base Learner Algorithms

A single base learning classifier is one of the machine learning techniques which follow the basic rules of machine learning. These techniques take the whole set of training data and apply only one of the ML algorithms to build a predictive model. As shown in Figure 3.3, the process of building the model is clearly implemented in a linear manner.

Below is a brief description of each of the single learner algorithms used in this thesis. The algorithms have been divided into several categories according to their functionality.

3.4.1.1. Function Based Approaches

- ✓ ***SimpleLinearRegression***[11]: SLR generates a regression model which minimises the sum of squared error or residual. It attempts to map the training

data points to the target variable through a linear relation. Formally, given training data X and their corresponding target values y, Linear Regression solves for a set of weights ***w*** that minimises the squared error on training data $(wX-y)^2$.

- ✓ ***LinearRegression***[11]: It is similar to the SLR algorithm defined above except in this algorithm the use of Akaike Information Criterion (AIC)[56] for model selection has been implemented. AIC is measuring the quality of each model relatively to others for giving dataset.

- ✓ ***MultilayerPerceptron***[11]: MLP is an implementation of Artificial Neural Networks (ANN) using Back-Propagation. MLP is used to modify the weight of hidden nodes based on their individual contributions to the final prediction. In addition, each of the nodes in the network uses a sigmoid function.

- ✓ ***SMOreg*** [57]: SMOreg is an implementation of a Support Vector Machine (SVM) for a regression task. Basic idea of support vector machines is to find optimal hyperplanes for linearly separable patterns. SVM uses different kernel functions to extend the patterns that are not linearly separable by transforming the original data into a new space.

3.4.1.2. Tree based approaches

DecisionTree takes the training set and represents the data as a tree structure (node and branches). A tree node represents a test of an attribute (question); each branch indicates an outcome of the test, and the last node, which is called the leaf, contains the class label. The constructed tree is used to classify any new data to its class. The new data will go through the tree from the first node (root) up to a leaf.

The root must hold the feature which best divides the dataset. Choosing the root is one of the critical issues in a DecisionTree design. A number of measurements are used to identify the best feature, for example Information Gain, Gain Ratio, and Gini index [11].

- ***M5P*** [58]: M5P is a tree-based regression approach. It builds a tree similar to in other tree-based models. However, its leaf node contains a regression model, rather than values, such as in other tree-based models. M5P normally generates smaller trees than regression trees and thus can be learned more efficiently and can handle higher dimensionality.

- ***DecisionStump*** [59]: This is a one level decision tree. The algorithm makes a decision from one value of the input feature. However, this algorithm is often used as a base classifier for other meta learning (ensemble learning) algorithms (see Section 3.2.2).

- ***RandomTree*** [60]: The RandomTree algorithm randomly samples the features at each node of the tree without performing pruning.

- ***REPTree*** [11]: REPTree is a regression tree that utilises information gain as the splitting criteria. The tree is pruned with reduced-error pruning and the values of numeric attributes are sorted once only.

3.4.1.3. Rule Based Approaches

- ***DecisionTable*** [61]: DecisionTable is one of the simplest classification approaches used in supervised learning. It is based on a matrix/table that contains features and instances from the training dataset. Therefore, the algorithm will take any new instance and search the entire table for a match (it could have multiple matches). If no matches are found, the best matching class of the dataset is returned. Otherwise, the best matching class out of the matching instances found is returned.

- ***M5Rules*** [62]: is a tree model that is similar to M5P (see above). The difference from M5P is that M5Rules creates a decision list for regression

problems using the *separate and conquer* [63] approach. In each iteration, an M5 Tree model is built, and the best leaf is selected to be made into a rule. The model produces a rule set that is as accurate, but smaller, than that of a tree based model (see Section 3.4.1.2).

3.4.1.4. Lazy Approaches

- ✓ ***Lazy.IBK*** [59]*:* IBK is a K-nearest-neighbour classifier. The algorithm tends to store the entire training sample before building the classifier. Once a new sample (data) is received, the algorithm will build the classifier. Hence, IBK uses a distance measure to locate the K closest instances to the new sample of data from within the training set. Then it uses these selected K instances to build the model. This algorithm can work for both regression and classification models.

- ✓ ***Lazy.LWL*** [64]: LWL is a Locally Weighted Learning algorithm. The weight is assigned using an instance-based method. Then the classifier model will be built from the weighted instances. This algorithm can carry out classifications using, for example, the Naive Bayes approach, or conduct regression by adopting linear regression.

- ✓ ***Lazy.KStar*** [65]***:*** KStar is a similarity-based model. For each instance to be predicted, it searches in the training set for the most similar instance. The prediction is determined by this most similar instance. It uses an entropy-based distance function, which makes it different from other similarity-based models.

3.4.2. Ensemble Learner Algorithms

Ensemble learning is an approach that uses multiple learning algorithms to build a single model. The idea is similar to a committee meeting, in which each of the members has an opinion on solving the matter they discuss. Different ideas can be

combined or a single thought can be agreed upon. In either case, at the end the chairman will call for a vote and the majority will win. This is exactly what happens when using ensemble techniques, but each has a different way of voting and taking the decision.

Another general definition, presented by [51], is as follows : "*Ensemble learning is a process that uses a set of models, each of them obtained by applying a learning process to a given problem. This set of models (ensemble) is integrated in some way to obtain the final prediction*."

The ensemble learning approaches are split into two main categories, namely homogenous (using a same base learning algorithm on different distributions) and heterogeneous (using different multiple learning algorithms). Learning algorithms in both categories aim to improve the performance of a model by reducing the variance and the bias of the dataset. Hence, an ensemble can be used to solve both classification and regression tasks [11]. The main focus of all ensemble methods (i.e. Bagging, Boosting) is to overcome the problems associated with weak predictors [66].

To obtain an ensemble learning method, three main steps have to be implemented, regardless of the type of the task [51]:

1. Ensemble Generation: this step is used to generate several samples, of which each will build a model using a single base learning algorithm.
2. Ensemble Pruning: this step will eliminate some of the models which have been generated in the first step. The purpose of this step is to reduce the size of the tree without affecting the accuracy. Ensemble pruning was introduced by [67].
3. Ensemble Integration: uses a voting or averaging strategy to combine the models and this strategy is used to predict any new cases.

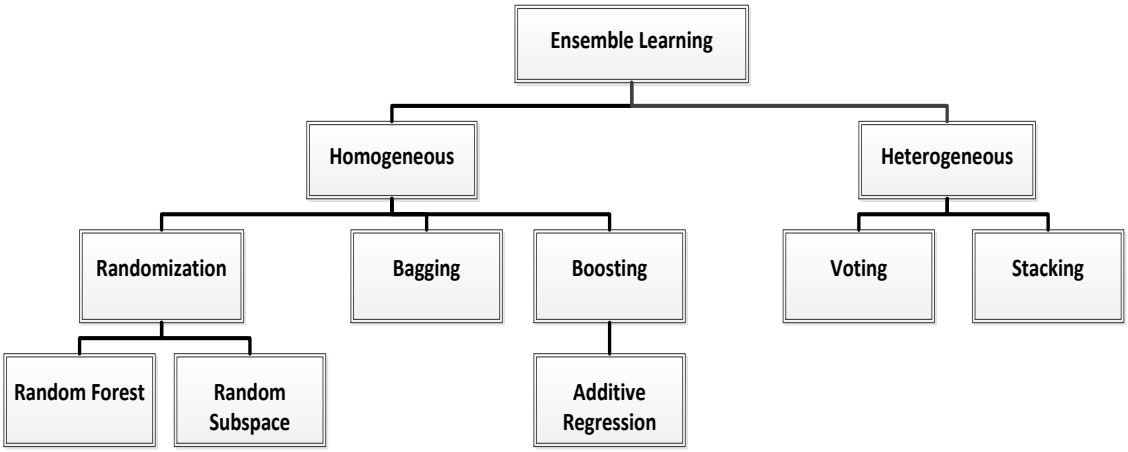

Figure 3.4: Ensemble Learning Hierarchy

Figure 3.4 demonstrates the hierarchy of ensemble learning classifiers. The following is a summary of the methods that is used in this thesis, with a graphical illustration of the basic homogeneous methods (see Figure 3.5):

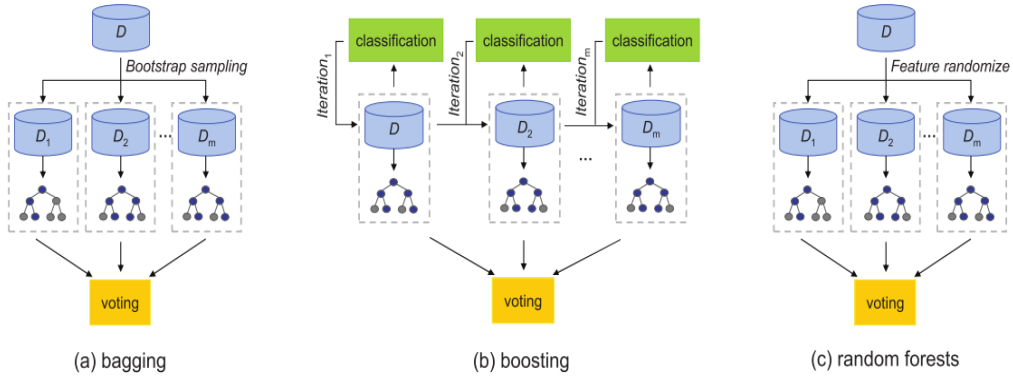

Figure 3.5:Fundamental concepts of the three basic homogeneous methods [42]

In the research presented in this thesis the most common ensemble learning algorithms, namely Bagging, Random Forest, Random Subspace, Addictive regression, Voting and Stacking are used to model the ground level ozone concentration. General explanations of each method with the view of

understanding their performance on the test dataset to be investigated are as follows:

3.4.2.1. Bagging[68]

Bootstrap aggregation (Bagging) is a common type of ensemble learning approach. It is based on the manipulation of a given training data set [44]. Bagging resamples the original input data by using the bootstrap method, randomly but with replacement (some data elements can be selected repeatedly, while others may not). The data included in each sample are different from each other; however, the sizes of these samples are equal. A separate classification model is developed from each sample using one single learning algorithm. Subsequently, the outputs of different models are integrated into a single prediction model. It uses either the weighted vote or the average vote, depending on the type of task (i.e. a classification task or regression task respectively).
Witten [11] stated that the ultimate model that results from Bagging often performs better than a single model that acts on the entire input dataset and never gets worse. The disadvantage of this method is that it does not work with a stable learning algorithm, e.g. a K-nearest-neighbour algorithm, where small changes do not affect the accuracy.

3.4.2.2. Random Forest [69][70]:

This is one of the Bagging tools. The Random Forest approach performs well in datasets which have more attributes than instances. Furthermore, the Random Forest approach performs better than Bagging due to the extra randomness present in the process of building the model. In other words, Random Forest is different to Bagging in the way that it splits the node of a tree. Instead of looking for the best point to split the node among the whole set of variables, Random Forest randomly picks sub-features to search for.

3.4.2.3. Random Subspace [71]

This is another tool of Bagging, which uses the same principle of sampling the training set. However, Random Subspace samples the training set based on features instead of the instances which Bagging uses. Random Subspace is more effective when the dataset contains a significant amount of duplication in the data features. In addition, it can operate efficiently when the dataset contains fewer instances than features.

3.4.2.4. Additive Regression [72]

Additive regression (Boosting) is a meta classifier that combines multiple linear regression models to enhance the performance. In each iteration, a model is created to fit the residuals from the previous model, in other words the current model usually depends on the performance of the previous model. During prediction, all outputs of the models will be added up to produce the final prediction result.

3.4.2.5. Voting [11]

Voting adopts the same mechanism used in Bagging except that it combines multiple models obtained using different learning algorithms to build the desired final model. The output of each classifier can then be averaged to produce the final model.

3.4.2.6. Stacking [73]

Stacking is an extended version of voting. It takes multiple classifiers which are trained using the original dataset. This process is called first level learning. Subsequently a new training dataset is produced from combining the output of

each individual classifier of the first stage to feed into a second level learning algorithm, named the meta classifier. The meta classifier is a single classifier which finds the best way to combine the outputs of first level to produce the optimal output.

3.5. K- fold Cross Validation [74]

K-fold cross validation is one of the several options that can be used to evaluate predictive models. This idea was initially introduced by Seymour Gerisser in 1993[75]. Subsequently a number of statistical researchers have used and further developed the theory of cross validation. Several studies such as [74] have proven that K-fold cross validation can reduce the error and provide a better approximations of generalization. However, it is computationally expensive as it trains and tests in every point.

In the proposed research the 10 fold cross validation has been used. The original dataset is randomly divided into K-different folds (K=10 equal size partitions, D_1, D_2,……D_K) where the training and testing process will perform K times. In each iteration 'i', the portion D_i will be held as the test set, while the remaining data portions will be used as the training set. At the end of each iteration, a model will be produced and the average of the prediction results obtained from the K different partitions, will be taken as the final result. This approach will reduce the variance of the model.

3.6. The Validation Metrics

In order to evaluate the accuracy of the prediction models, four objective metrics have been used, namely the Correlation Coefficient(CC), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Relative Absolute Error(RAE). These are metrics popularly used to compare accuracy in modelling O_3 (i.e. a single parameter) concentrations. All of the matrices are aimed to calculate distance between the estimated value and the actual/true value. The four metrics can be defined as follows:

3.6.1. Correlation Coefficient (CC)

Determines the linear relationship between input variables (X) and target variables (Y). It takes values between -1 and 1. The Correlation Coefficient is defined as follows:

$$CC_{X,Y} = \frac{E\,[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

where σ_X and μ_X are respectively the standard deviation and the mean of X, while E represents the expectation. In addition, σ_Y and μ_Y are defined similarly for Y. A positive value of Correlation Coefficient means that the two variables move in the same direction with respect to their means. A negative value means they move in opposite directions with respect to their means. A value close to 0 means the two variables have little linear dependency (see Figure 3.6). This means, for the predictions of the proposed work in this thesis, the Correlation Coefficient should be maintained close to 1 as much as possible, as this would facilitate training accurate models.

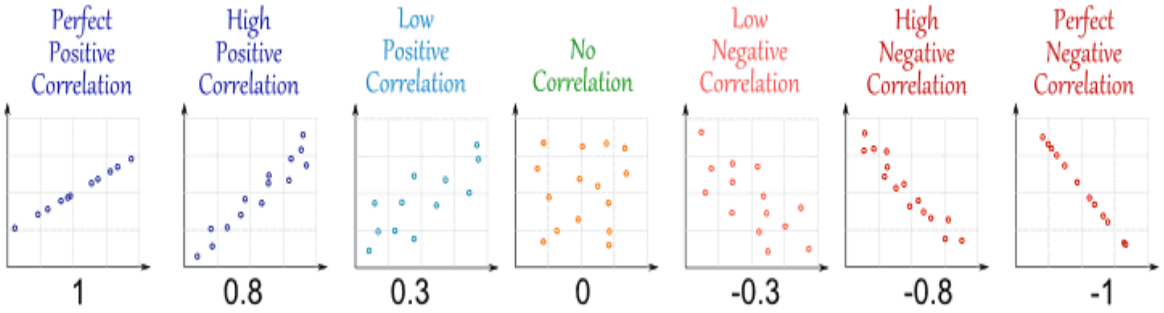

Figure 3.6: Liner correlation : the interpretation of different values [76]

3.6.2. Mean Absolute Error (MAE)

MAE refers to the sum of individual absolute errors normalised by the number of samples. The individual error is defined by the difference between ground truth and predicted value for a sample. Mean Absolute Error is given by the following equation.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - y_i|$$

where p_i is the predicted value, y_i is the ground truth, n is the total number of samples and | | is the notation used to calculate the absolute value of a term within. This kind of measurement is more tolerant to large errors, the reason is because the error is not squared compared to for e.g., RMSE [77].

3.6.3. Root Mean Squared Error (RMSE)

Root Mean Squared Error is a modification of the mean absolute value, with the absolute value of an individual error term replaced with a square. Its definition is given as follows:

$$RMAE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - y_i)^2}$$

MAE and RMSE are both measuring the average difference between the predicated and actual value [78]. However, RMSE is the most commonly used to measure the goodness-of-fit. RMSE gives more attention to the large errors due to its square term [77].

3.6.4. Relative Absolute Error (RAE)

RAE is a calculation of the variance of a model where the units are not important when comparing between models. The previous two measurements (MAE and RMSE) depend on the scale of data, and thus RAE can be very helpful when comparing different data with different scale.

$$RAE = \frac{\sum_{i=1}^{n}|p_i - y_i|}{\sum_{i=1}^{n}|\bar{y} - y_i|}$$

Where $\bar{y}$ is the mean value of y.

3.7. Filters and Optimisers in WEKA

WEKA toolkit provides several filters for different proposes that can be used to optimise the results obtained by a selected machine learning algorithm. The research proposed in this thesis has used some of these filters to help improve the accuracy and the performance of the models generated. Two types of filters have been used and are thus described below. They are, feature/attribute selection filters and parameter based learning algorithm optimising filters:

3.7.1. Feature Selection Filters [11][79]

Feature selection or attribute selection plays an important role in removing redundant features from being used in creating a model. Hence only the most relevant features will be selected for the purpose of modelling. This results in many advantages that are described below. It is noted that feature selection filters operate differently to wrapper filters as they evaluate the significance of attributes independently of any learning algorithms by focusing on general data characteristics and relationships.

The advantages of performing attribute selection on data are:

- **Reduces Overfitting**: Less redundant data means less opportunity to make decisions based on noise.
- **Improves Accuracy**: Less misleading data means that the modelling accuracy may improve.
- **Reduces Training Time**: Less number of attributes means that the amount of data used in the model generation will be reduced, which means that the algorithms can train faster.

They are several attribute selection filters which can work in conjunction with either classification tasks, regression tasks or both types. These filters usually use different methods to come up with the final decision on selected attributes [80].

There are two categories of filters implemented within the software package which employed in this thesis. Filters, (1) which evaluate a subset of the attributes and those (2) which evaluate individual attributes. Each of the attribute selection filters use an attribute evaluator and a search method that follows. Following are general explanations of Attribute Evaluators and Search Methods implemented and this used in the proposed research.

1) **Attribute Evaluator**: This filter weighs each subset of attributes and assigns a numeric value, which monitors the search process. The evaluations could be measured by building a model and assessing the accuracy of the model. Some of attribute evaluation methods are listed below:

   - **CfsSubsetEval**: Individually evaluates the ability of each attribute and assign high scores to the features that are highly correlated to the class attribute, but have low correlation to the other features in the dataset.
   - **ReliefFAttributeEval:** This filter is used with the ranker search method which generates a rank list. It randomly picks an instance from the original dataset and then finds its nearest neighbours, from the same and opposite classes. The values of attributes for each of the instances (sampled and the nearest neighbours) will be compared and the score of each attribute will be updated relatively.
   - **Principal Component Analysis (PCA):** This is a filter which transfers the data by changing the attributes and combining them into a new form. This filter usually requires the use of Ranker search method.

2) **Search Method:** The search method refers to the method in which the attribute selector searches through a possible subset of features. There are different approaches to search namely, BestFirst, GreedyStepWise, and Ranker methods. Each method uses a different approach to building the list of the attributes and provides a weight for each attribute.

3.7.2. Parameter Based Llearning Algorithm Optimisers

All learning algorithms consist of a number of parameters and often the selection of the values of these parameters that will ensure that the learning algorithm will work optimally is a challenge.

In the proposed research one of the parameter based optimisation filters, "CVParameterSelection" has been utilised to find the optimal parameters for the selected classifier. However, this type of the filter works only for the single base learner (i.e. it cannot work for meta learner classifier) algorithms and cannot thus be used for optimising ensemble learning algorithms. The algorithm requires the user to provide with ranges for each of the parameters associated with a learning algorithm. Then the filter attempts to find the optimal setting of the parameters by trying out model creation using several combinations of parameters within the range specified. The software package utilized, also offers other types of parameters optimisation algorithms namely Grid search, MultiSearch and Auto-WEKA. Details of each of these methods are presented in [81]. Within the research context of this thesis the method CVParameterSelection has been used, due to its simplicity and ease of use.

3.8. Summary

This chapter has presented detailed information about the background of systems, methods and algorithms used within this thesis to support the research conducted in the modelling, prediction and forecasting of ground level ozone concentrations. The chapter presented the apparatus used to measure the various concentrations of gases considered in the modelling of ground level ozone, the machine learning algorithms used for learning and creating the models, the filters used to pre-process data, optimise the learning algorithms and also the metrics used in the evaluation of the performance of the machine learning algorithms.

It is assumed that the vital background information presented in this chapter will help readers better understand the usage of various apparatus, techniques, algorithms and

systems, proposed in the contributory chapters of this thesis, i.e. Chapters 5, 6, and 7 without having to refer to additional literature. However, readers interested in more details are referred to the original publications and references where appropriate.

CHAPTER 4

Data Collection and Representation

The experiments conducted within the research scope of this thesis were based on two datasets (i.e. regression data), one obtained from Sohar University, Oman and the other obtained from Department for Environment Food & Rural Affairs (DEFRA), UK. The dataset from Sohar University was collected over a short period of time and is complemented by the presence of numerous environmental parameters that are required for the modelling of ground level ozone concentrations. However, this dataset consists of many data recording errors and missing values (clear identified errors). The dataset obtained from DEFRA, UK, has been recorded over a longer period of time and is ideal for time series prediction and/or forecasting of ground level ozone.

With the two datasets used in the experiments, two different techniques were employed for missing value and outlier removal. Since the Sohar data set was not used for time dependent series analysis the cleaning process employed removed the whole record, as reducing the number of records do not have a significant impact on obtainable model accuracy. However when using the DEFRA dataset as it was used for time series analyses in which the number of records used has an impact on the obtainable model accuracy, the imputation method [82] was used to fill missing values.

For clarity of presentation this chapter is divided into three sections. Section 4.1 introduces the Sohar University dataset and its preparation process for modelling and prediction of ground level ozone. Section 4.2, presents the DEFRA dataset and the pre-processing techniques used for its preparation for modelling and prediction of ground level ozone. Finally, section 4.4 concludes the chapter providing a summary.

4.1 Sohar University Dataset

4.1.1 The Sampling Site and Data Gathering

Measurements were recorded across the Sohar Highway (SHW), Oman, in front of the main entrance to the Sohar University (SU) with a Differential Optical Absorption System (DOAS) instrument (see Section 3.2) that was professionally installed (see Figure 4.1 for an aerial view of the installation of the system). The light beam travels a round-trip of 477 meters from A, which is located on the roof of a main administrative office building of SU, to B, where a reflector (or receiver) is installed on the top of a mosque minaret, as illustrated in Figure 4.1. The SHW has two lanes in each direction and an additional two single carriageway roads, in parallel, on both sides, bringing the total number of lanes to eight. Additionally, there is the SU car park, marked as C, where vehicular traffic may be present and thus may result in higher levels of ground level O_3 concentrations. The reflected light beam across SHW is captured at A and transferred by an optical fibre to the DOAS instrument, where a spectrometer splits the light into narrow wavelength bands using an optical grating. Subsequently, these bands are processed and evaluated to obtain the best estimation of the concentration of the monitored gases in the light path. In order to capture the rapid variations of the concentrations of gases present in the space of the monitoring path, evaluations of light captured by the DOAS instrument is performed every 30 seconds for the measurement of the concentrations of O_3, NO_2, and SO_2 gases and every one minute for measurement of the concentration of BTX. Additionally, the meteorological parameters, including wind speed and direction, relative humidity, pressure, temperature, precipitation, global solar radiation etc., are separately measured by sensors located on the roof of the SU building at A. The height from ground level was approximately 12 metres.

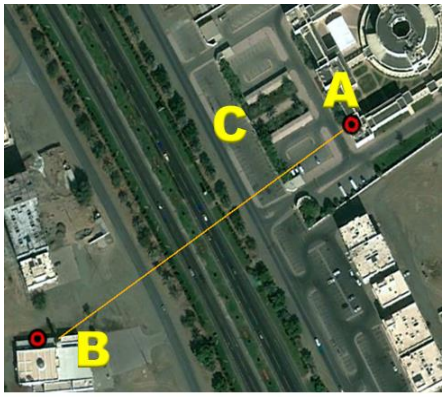

Figure 4.1: Sampling path of the DOAS instrument installed on the premises of Sohar University, Oman; A = light emitter location, B = reflector location and C = car park

The dataset used for the experiments in Chapters 5 and 6 were captured by the Sohar University DOAS system during 2013/2014 at a sampling rate of one hour. However, due to a technical fault in the system, the dataset collected during the specified period is not continuous. Nonetheless, a sufficiently large dataset was gathered to make the experiments conducted in Chapters 5 and 6, statistically relevant.

The Sohar University dataset contains a total of 6,744 instances, spread across the years 2013-2014, as detailed in Table 4.1. Note that there is a substantial data collection gap between 23rd of August 2013 to the 2nd of March 2014, during which time the DOAS system was non-operational due to an essential maintenance repair.

Table 4.1: Sohar University, Dataset Description

Dataset	2013			2014			Total number of records
	Start Date	End Date	No. of Rec.	Start Date	End Date	No. of Rec.	
	1st April 2013	23rd Aug. 2013	3480	1st March 2014	14th July 2014	3264	6744

4.1.2 Dataset Representation

The Sohar University dataset is a sequence of measurements presented in a time series. The measurements include concentration values of eight gases measured in µgm^{-3} and readings of six environmental parameters. Table 4.2 lists the 14 attributes of each measured data value with their descriptive statistics.

Table 4.2: Attributes of the Sohar University Dataset

2013-2014	Unit	Min	Max	Standard deviation	Mean
Sulphur Dioxide (SO_2)	μgm^{-3}	1.61	15.11	2.33	4.96
Nitrogen Dioxide (NO_2)	μgm^{-3}	0.02	83.99	16.65	18.24
Ozone (O_3)	μgm^{-3}	0.85	139.50	24.25	43.25
Benzene (C_6H_6)	μgm^{-3}	0.05	19.56	4.17	6.13
Toluene (C_7H_8)	μgm^{-3}	0.73	47.14	7.77	15.16
p-Xylene (C_8H_{10}(p))	μgm^{-3}	0.10	8.75	1.18	3.30
m-Xylene (C_8H_{10}(m))	μgm^{-3}	0.69	5.44	0.52	2.44
o-Xylene (C_8H_{10}(o))	μgm^{-3}	0.80	58.15	6.91	29.56
Temperature	ºC	16.19	45.06	3.53	31.10
Relative Humidity	%	8.47	93.57	19.33	64.38
Pressure	kPa	98.94	102.89	0.56	100.19
Global Radiation	W/m^2	-2.75	1120.24	247.95	201.13
Wind speed	m/s	0.31	6.266	1.02	1.77
Wind Direction	degree	0.11	359.99	91.50	137.52

4.1.3 Data Pre-processing

The original dataset was subjected to two data pre-processing operations, i.e., removal of missing values and outliers, and data transformations.

4.1.3.1 Removal of Missing Values and Outliers:

Data cleaning operations listed under preprocessing algorithms previously were utilised for the removal of missing values and outliers. The two filters interquartileRange (filters -> unsupervised -> instances -> interquartileRange) and removeWithValues (filters -> unsupervised -> instances -> removeWithValues) were used to clean the input raw data recorded by the DOAS system. Once the data is cleaned as above the dataset should typically be ready for modelling.

4.1.3.2 Data Transformations:

Since the wind direction is originally measured as an angle from the north in a clockwise direction, with values ranging from 0-360 degrees, the originally recorded wind related data will have to be re-represented to avoid 0 and 360 degree directions being considered as different. In order to deal with this issue, the Wind Speed (WS) and Wind Direction (WD) have been combined and divided into two orthogonal compenents,

$$u = WS \cos(WD) \qquad (4.1)$$

$$v = WS \sin(WD)$$

The (u,v) parameters replace (WS, WD) in the modelling process.

4.2 DEFRA Dataset

The dataset used for the purpose of time series analysis and forecasting of ground level ozone concentrations (see Chapter 7) was obtained from the Department for Environment Food & Rural Affairs (DEFRA), UK, available at https://uk-air.defra.gov.uk/. It contains data from approximately 300 environment monitoring sites spread cross the UK aimed at monitoring air quality. The stations are organised

into a network of sites which gather air quality and other environment related information mostly using different methods. Each networked station records different pollutant parameters, which mainly depends on the purpose of the monitoring station and the equipment/methods used for data gathering. Hourly measured ozone concentrations are usually monitored and recorded at each station.

For the studies carried out in Chapter 7 data gathered in a monitoring station in London, London Marylebone Road, was selected using a random selection process if the station to be scrutinised. This data consists of continuous records from August 2010 to April 2016. Each record consists of concentrations of six gases, namely ozone (O_3), nitrogen oxide (NO), nitrogen dioxide (NO_2), nitrogen oxides (NO_x) sulphur dioxide (SO_2), and carbon monoxide (CO), and three meteorological parameters namely wind direction (WD), wind speed (WS) and temperature (T). The total number of records collected for the said period was 50208 (see Table 4.3).

Table 4.3: Statistical presentation of DEFRA dataset

Attribute	Missing values(h)	Missing values in Percentage	Min	Max	Mean	StdDev
CO	1508	3%	-0.291	5.782	0.571	0.316
No	1240	2%	-0.031	852.342	138.66	114.897
NO_2	1240	2%	4	304	91.537	43.095
NO_x	1240	2%	8	1466.964	303.944	214.156
WD	1797	4%	0	360	197.54	96.903
WS	1797	4%	0	13.1	3.678	1.756
T	1797	4%	-10.4	32.8	10.078	6.098
O_3	5010	10%	-0.93	492.14	15.8	14.607
SO_2	3602	7%	-0.865	48	7.768	5.534
Total number of records	50208					

4.2.1 Data Pre-processing

The original set of 50208 data instances require cleaning in the form of removal of missing values and outliers. Further some data features require transformation, prior

to the data being used for modelling and prediction of ground level ozone concentrations.

4.2.1.1 Missing Value and Outlier Removal

The Imputation Method [82] was used for filling missing values. The imputation method fills missing values via the application of interpolation algorithms, based on data captured at the same time, the day before and after. In the experiments conducted in this research by adopting the above data cleaning strategy, the number of data instances that were finally available for modelling, became the originally expected number of records. It is noted that the number of data instances sued in modelling plays a key part in the model accuracy.

4.2.1.2 Data Transformations

For the reasons described in Section 4.1.3.2, the Wind Speed (WS) and Wind Direction (WD) have been combined and divided into two orthogonal compenents as in Equation 4.1 (refer to Section 4.1.3))

4.3 Summary

The machine learning based approaches proposed for the prediction of ground level ozone both spatially (without the consideration of time, but with respect to attributes known to create ozone) in Chapters 5-6 and temporarily (forecasting with respect to time) in Chapter 7 have used two different datasets. Whilst the former uses a dataset recorded during a short period of time in the city of Sohar, Oman, the latter uses a dataset recovered from a particular area of London, UK, provided by the DEFRA, UK. A close analysis of both datasets showed that the data needed cleaning and then some mathematical transformations to ensure that they can be effectively used in constructing accurate prediction and forecasting models. This chapter has proposed the methods that are needed for the removal of missing values and outliers and also proposed appropriate data transformations for certain given attributes.

CHAPTER 5

Modelling Atmospheric Ozone Concentration Levels

The aim of this chapter is to investigate the performance of meta learning ensemble algorithms in the prediction of ground level ozone (O_3) based on the concentration of other atmospheric gases and meteorological parameters that have an impact on the formation of ground level ozone. It is noted that in this chapter, the dependence of ozone concentration on time, is not considered. Nevertheless, time dependence is investigated in detail in Chapter 7.

5.1 Introduction

Most of the historical and recent studies in air pollution modelling, monitoring and analysis have employed standard non-linear classifiers, mainly either Artificial Neural Networks (ANN) or Support Vector Machines (SVM) to model the atmospheric ozone concentration based on supervised learning. Some recent studies in environmental modelling have analysed the use of a broad variety of learning algorithms, but such studies have been limited to analysing data with standard software packages, using the default settings of model parameters. In application areas outside environmental pollution monitoring, some attempts have been made on using ensemble learning algorithms for data modelling and prediction giving improved prediction results. However, even these studies have been limited by the number of different algorithms investigated and the constraints under which they have been applied. To our best knowledge ensemble methods have not been applied and investigated comprehensively in the prediction of O_3. Therefore, in this chapter a comprehensive investigation is carried out on the performance of three different meta learning ensemble approaches, namely, Bagging, Voting, and Stacking to build models for the prediction of ozone concentration in the city of Sohar, Oman. Moreover, the results of ensemble learners are compared with the performance results of a significant number of popular learning algorithms used in the literature and investigated within the research context of thesis and presented in Chapter 3.

The prediction of ground level ozone concentration is based on the concentrations of seven gases (nitrogen dioxide (NO_2), sulphur dioxide (SO_2), and BTX (benzene, toluene, o-,m-,p-xylene) and six meteorological parameters (ambient temperature (T), air pressure (P), wind speed (WS), wind direction (WD), solar radiation (SR), and relative humidity (RH)).

The dataset considered in this work was obtained from Sohar University, Oman, which used a DOAS instrument [54] (see Chapter 3) to gather environmental data within the premises of the university campus, in the city of Sohar, Oman. The dataset includes concentrations of eight gases and six meteorological parameters as defined above.

The modelling results presented in this chapter show an impressive prediction performance improvement obtainable by using meta learning algorithms (i.e., Bagging, Voting, and Stacking) as compared to the traditionally used learning algorithms. The performance accuracy of these different meta learning methods were approximately the same giving an average of 0.91 correlation coefficient in prediction accuracy though they demonstrated a significant increased accuracy over the traditional methods.

For clarity of presentation the chapter has been divided into several sections. Section 5.2 presents the motivation behind the proposed research. Section 5.3 provides a summary of the experimental methodology and Section 5.4 presents details of experimental settings adopted. Section 5.5 presents the design details of the various experiments conducted to compare the prediction accuracy of ground level ozone when different ensemble learning algorithms are used. Section 5.6 presents the experimental results and a detailed analysis of the results. Finally, Section 5.7 summarises the investigations conducted and make conclusions based on the analysis presented in Section 5.6.

5.2 Motivation

A number of studies in the field of environmental science and engineering have focused their interest on constructing models to predict the concentrations of gases that result in air pollution. The majority of environmental researchers tend to use Artificial Neural Networks (ANN) and Support Vector machines (SVM) to predict

ozone concentration[20],[24],[26],[27]. Although there are more advanced and recent data mining / machine learning techniques, such as Ensemble Learning approaches [11], only few attempts have investigated their use in predicting atmospheric or ground level ozone concentration[8]–[10]. However the investigations of [8]–[10] were limited in the fact that only the ensemble classifier Bagging was used adopting only the default single classifier RepTree as the base classifier of the ensemble algorithm, as implemented within the data mining software package, WEKA. Our detailed investigations revealed that in the field of air pollution monitoring, no attempt has been made to test other ensemble classifiers, select the best base classifier or to optimise the performance of the base classifier of the ensemble of classifiers based on various possible parameter selections, all of which can lead to significant improvements in prediction accuracies. On the other hand, several attempts have been made in areas beyond air quality prediction in the use of ensemble classifiers, such as in bioinformatics, medicine and marketing, to build more efficient predictive models [83]–[86]. This work has shown that ensemble classifiers outperform the corresponding single classifiers and that the ultimate answer to the question, which classifier works best, depends on the dataset. It is clear that different datasets, in particular from different application domains, are statistically different and this has a high impact on the variability of results obtainable from different classifiers.

From the review of literature conducted and summarised above, a lack of research into effectively utilising Ensemble learning algorithms to predict ozone concentration was identified. Therefore, the research proposed in this chapter aims to find accurate models that can be used to predict ground level ozone concentrations, given a multitude of environmental parameters and the concentrations of gasses that are known to result in the creation of ozone. An investigation is carried out comparing the performance of several machine learning techniques. Multiple predictive models were built using popular single classifiers used for regression (e.g. Multilayer Perceptron (MLP) and Support Vector Machines) and selected ensemble learning algorithms (homogeneous and heterogeneous), (refer to Figure 5.1). In all experiments the data mining software tool, WEKA (Waikato Environment for Knowledge Analysis) is used as implementations of the learning algorithms.

5.3 Proposed Approach

The proposed approach adopts standard data mining procedure that involves data pre-processing prior to data modelling using machine learning. WEKA (version 3.7.11) is a toolkit that supports open source software implementation and operation of a large number of options for both data pre-processing and modelling.

The aim of this section is the analysis of the performance of the most efficient ensemble learning algorithms in the prediction of O_3. For this purpose, the author has compared the performance of ensemble learning algorithms with the performance of those single classifiers that have been popularly used in literature. It is noted that the focus of this chapter is to investigate the use of the most popular homogeneous approach, Bagging and the two heterogeneous approaches, Voting and Stacking (refer to Chapter 3 for illustration of different classification of ensemble learning algorithms).

In order to evaluate the performance of the Ensemble Learning Algorithms and compare them with single learning algorithms, two key sets of experiments were designed, implemented and tested. They are detailed in the following sections.

Initially, training phases based on different classification algorithms for predicting O_3 concentration were performed. Subsequently, the prediction performance of different algorithms, were examined using ten-fold cross validation (see Section 3.5). Various evaluation metrics have been utilised to analyse the results. It should be noted the key focus of the research conducted is not time-series analysis of O_3 concentration (i.e. predicting how O_3 concentration changes with time) but how to predict O_3 concentration based on the concentrations of the primary pollutant gases and the environmental parameters that are likely have an impact. In particular, when O_3 creation is assumed to be due to the production of primary pollutant nitrogen dioxide, generated by vehicular traffic in this area, the time dependent analysis is not essentially useful.

Since the experts have proven that there is no single machine learning algorithm can be applicable to all types of data, the first group of experiments was implemented by adopting all the applicable algorithms provided in WEKA.

5.4 Experiment Settings

Two WEKA working environments were used, i.e., the Explorer and the Experimenter [12], each used for different purposes. In our experiments, 16 machine learning algorithms implemented within WEKA have been used for testing and performance comparison, with their default parameter settings being used and using ten-fold cross validation to evaluate individual models (Table 5.2 lists all the classifiers). Furthermore, rigorous studies of the use of meta learning classifiers (Bagging, Stacking, and Voting) were carried out by testing it with different base classifiers. The accuracy of the model was evaluated using four widely used evaluation measurements: Correlation Coefficient (CC), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Relative Absolute Error (RAE) (see Chapter 3).

In the data captured by the DOAS (see Section 3.2), missing values are recorded as -999. A careful analysis of the captured data also revealed that there are also measurement outliers, which would have resulted from temporary sensor malfunctioning instances due to dust, high temperatures and overheating. Therefore, the data has been pre-processed (outlier removal and data transformation) using the procedure explained in Chapter 4. After the data cleaning procedure, only approximately 62% (4,173 out of 6,744 instances) of the original dataset were utilised in the modelling phase.

5.5 Experimental Design

The experiments are conducted and presented in two groups for the purpose of clarity.

Group 1: Investigation on the use of single learner algorithms vs homogenous ensemble learning algorithms (Bagging, Random Forest, Random Subspace, and Additive Regression) for the prediction of ground level ozone.

The experiments in this group were broadly divided into two categories. In the first category sixteen models were constructed to include thirteen single base classifiers and the homogeneous ensemble classifiers, Random Forest, Random Subspace and Additive Regression. In the second category, all of the models in Category 1 were used as the base classifier of a Bagging meta classifier. In each category, the

performance of the above mentioned sixteen different learning algorithms were investigated, divided into five different algorithm categorisations, depending on their functionality: These include Function based (Linear Regression, Multilayer Perceptron, Simple Linear Regression, and SMOreg), Lazy approaches (IBK, Kstar, and LWL), meta classifiers (Additive Regression, and Random Subspace), Rule based (DecisionTable, and M5Rule) and Tree based (Decision Stump ,M5P, Random Forest, Random Tree, and REPTree) classifiers. The prediction accuracy results of each model were evaluated using the metrics mentioned in Section 5.4, (See Table 5.2).

It is noted that the experiments in Group 1 is divided into two sub-groups for clarity of presentation, according to the corresponding WEKA experimental environment being used (i.e. WEKA's Explorer and Experimenter working environments). Table 5.1 summarises the two sub-groups, which are numbered accordingly for clarity of referencing.

Table 5.1 : Group 1 Experiments Description

Experiment	**Description**	**WEKA Environment**
Division 1	Compare the accuracy of 16 classifiers, individually and as base classifiers in Bagging ensemble learning	Explorer
Division 2	Evaluate multiple models resulting from different classifiers	Experimenter

Group 2: Investigation of the use of heterogeneous ensemble learning algorithms (Voting and Stacking) for the prediction of ground level ozone.

Stacking and Voting have been categorised as heterogeneous ensemble learning approaches as both methods make use of multiple single base classifiers. In particular Stacking has an extended layer. This extra layer (named Layer 1) is fed with the output of the first layer and uses a single base classifier to produce the desired final model.

In this group the experiments were designed and categorised to compare the performance of the various models described above, under different paradigms. The experiments were broadly divided into two sub-groups, namely, those that use Voting and those that use Stacking. For Voting and the Layer 0 of Stacking the author have used the same combination of single base learning algorithms. For the Layer 1 of stacking a number of additional single base learning algorithms have been used.

In order to rigorously compare the performance of different combinations of single base classifiers within heterogeneous ensemble learning algorithms, the following criteria were used in selecting the constituent single base classifiers of Layer 0 (Refer to Table 5.4):

- All three experiments used either 3 or 5 algorithms as they were amongst the most accurate, when used as single base learning algorithms as demonstrated by the Group 1 experiments. Note that in Experiment 2, all three algorithms consist of tree base learning algorithms (**Experiments 1, 2, and 3)**.
- This experiment repeats Experiment 2, with the addition of two further algorithms from the tree base category of learning algorithms. The aim is to determine the effect of adding a further number of learning algorithms of the same type on the ensemble's accuracy (**Experiment 4)**.
- This experiment contains one classifier from each separate category of learning algorithms as defined in WEKA software (**Experiment 5 and 6)**.
- Contains a combination of algorithms that resulted in the worst performance accuracy, when tested as single base classifiers in Group 1 (**Experiment 7)**.
- Contains a combination of algorithms that resulted in the best and worst performance accuracy, when tested as single base classifiers in Group 1 **(Experiment 8)**.

As Stacking contains an additional layer of learning algorithms, termed as the Layer 1 classifiers, our experiments used the single base learning algorithm that performed the best, i.e. Random Forest, the one which is algorithmically simplest, i.e. Linear Regression and few further selected classifiers, have been used as the Layer 1 classifier.

For improved clarity,

Figure 5.1 illustrates a block diagram that explains the general structure of experiments designed and presented above, within Group 1 and Group 2 of experiments.

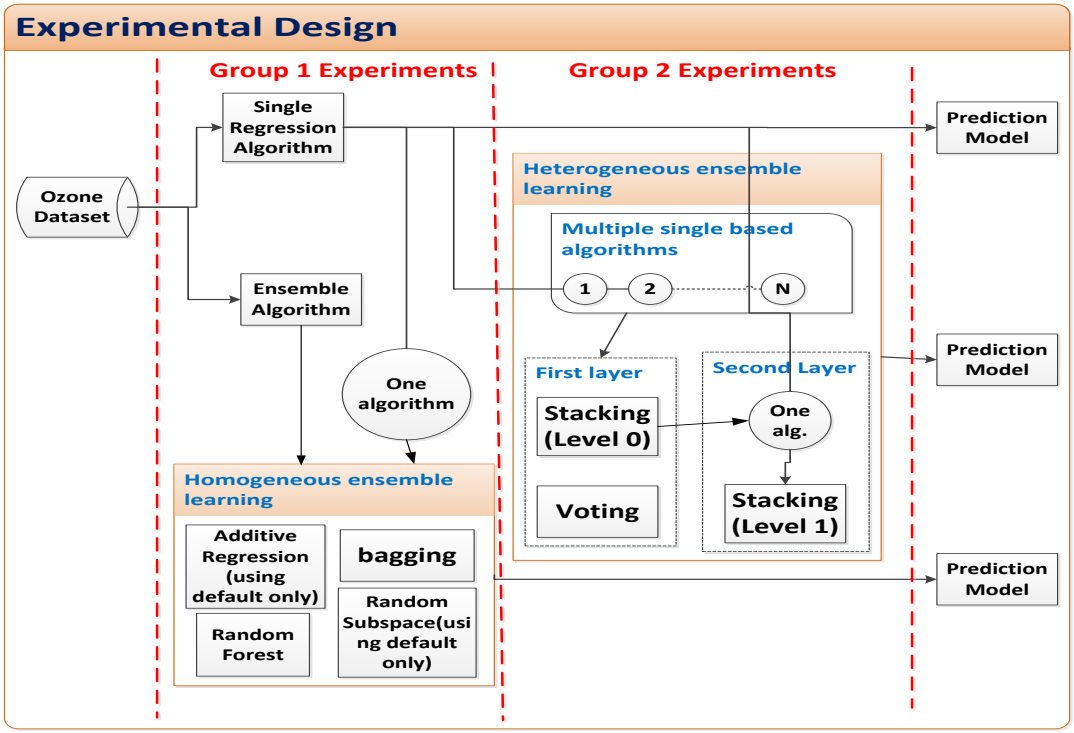

Figure 5.1: Categorisation of Experimental Designs

5.6 Results and Discussion

The results can be presented and analysed within the two experimental groups, as follows.

5.6.1 Group 1

Table 5.2 tabulates the results of the experiments carried out within Group 1. The purpose of the experiments within this group was to comprehensively compare the performance of various single base learning algorithms against those of four

homogeneous ensemble learning algorithms, namely Random Forest, Random Subspace, Bagging and Additive Regression.

Table 5.2: Results of Group 1 Experiments

Algorithm		Section 1: Without Bagging				Section 2: Using Bagging			
		CC	MAE	RMSE	RAE	CC	MAE	RMSE	RAE
Ensemble Classifier	Random Forest	0.91	7.52	10.32	40.16 %	0.92	7.08	9.75	37.81%
	Random Subspace with REPTree	0.90	8.15	11.29	43.50 %	0.91	7.71	10.66	41.15%
	Additive Regression	0.81	10.96	14.24	58.53 %	0.84	10.04	13.26	53.63%
Single Base Classifier	M5Rules	0.89	8.15	11.29	43.52 %	0.89	8.05	11.07	43.01%
	M5P	0.89	7.92	11.02	42.28 %	0.90	7.70	10.62	41.12%
	REPTree	0.86	8.96	12.48	47.84%	0.90	7.52	10.41	40.16%
	Multilayer Perceptron	0.85	9.81	12.95	52.38 %	0.90	7.64	10.45	40.79%
	Lazy.IBK	0.85	9.23	13.16	49.30 %	0.89	7.92	11.27	42.29%
	Lazy. KStar	0.86	8.56	12.50	45.70 %	0.87	8.14	11.86	43.47%
	Linear Regression	0.84	9.67	13.22	51.65 %	0.84	9.68	13.22	51.68%
	SMOreg	0.84	9.51	13.38	50.80 %	0.84	9.51	13.38	50.79%
	Random Tree	0.82	10.40	14.71	55.54 %	0.91	7.47	10.20	39.89%
	Decision Table	0.79	10.95	14.91	58.45 %	0.78	11.17	15.65	59.62%
	Lazy.LWL	0.69	13.48	17.70	71.98 %	0.73	12.88	17.03	68.8%
	Simple Linear Regression	0.58	14.89	19.79	79.49 %	0.63	14.27	18.98	76.18%
	Decision Stump	0.55	15.60	20.28	83.32 %	0.59	15.04	19.67	80.30%

It can be observed that there is a significant improvement in the prediction accuracy of O_3 concentration, when homogeneous ensemble learning algorithms are adopted. In particular, comparing the experiments of Section 1 with those of Section 2 reveals the ability of Bagging to improve prediction accuracy, as depicted by a noticeable reduction in the MAE and RMSE (see Table 5.2), as compared to using the relevant single base classifier by itself.

In addition, the accuracy of prediction performance of homogeneous ensemble learning algorithms (with the exception of Additive Regression), offer better

prediction accuracy when compared with performance of the single base learning algorithms widely used in literature.

It is further noted that when Bagging is used with Additive Regression as the base classifier, the performance is improved as compared to when Additive Regression is used on its own. This illustrates the superior and reliable performance of the ensemble algorithm adopted within Bagging. The performance accuracy of Additive Regression as compared directly with Random Forest and Random Subspace, is worse. i.e. a model that ensembles different simple Linear Regression models as against the two, more accurate tree based, non-linear models, Random Forest and Random Subspace.

Overall when the prediction accuracies of Group 1 experiments are considered (see Table 5.2), the best overall performance is indicated by Random Forest, either used as homogenous ensemble classifier on its own (CC=0.91, compare results of Section 1 of Table 5.2) or as the base classifier of the ensemble learning algorithm, Bagging (CC=0.92, compare results of Section 2 of Table 5.2). Moreover, the experiments in Group 1 further revealed that bagged Random Subspace and Random Tree performed as accurate as Random Forest, when used independently of Bagging. This result confirms the conclusion of [87] who showed that combining Bagging and Random Subspace has a comparable performance to Random Forest.

In the literature, the Multilayer Perceptron (MLP) (i.e. Artificial Neural Networks) is the most common learning algorithm used to predict atmospheric ozone concentration. According to [88], Bagging can be used as a solution to the local minima related and the over fitting problems from which MLP suffers. Therefore, applying ensemble methods such as Bagging to an MLP should enhance the accuracy of the MLP in general. Table 5.2 compares the accuracy of the MLP when used as a single base learning algorithm to the performance of the bagged MLP and reveals a 5% increase in accuracy.

It is observed that the SMOreg is the only base classifier that was not been affected when Bagging was applied, with the Correlation Coefficient remaining unchanged at 0.84. This similarity is due to the stability of the SVM algorithms[89],[90]. However, [91] shows that bagged SVM can perform better

for some dataset, although it can give equivalent results to single SVM for other dataset.

In addition to the above experiments, evaluations of the predicted models results were further examined, to evaluate the performance of multiple classifiers, which the Explorer environment cannot provide. The results of the previous experiments did not provide the statistical significance of the improvements. Therefore, WEKA's Experimenter environment was utilised to obtain this additional information. A statistical test (Paired T-Tester [92] corrected) was used to calculate the statistical significance between the different predictive models. The performance of the classifiers were examined using 10 fold cross validation and were compared using the Correlation Coefficient. In addition, the confidence interval between the classifiers was set to 5% (a default setting). Selected algorithms used for experiments in Group 1 (see Table 5.2) were examined in this experiment. The focus was on comparing the influence of ensemble learning on the most widely used classifiers in the study area (MLP, SVM). Table 5.3 demonstrates the three experiments implemented for the evaluation.

The results of the experiments are shown in Table 5.3. Note that the characters, v or *, appears beside the results to indicate the level of significance. Since the first classifier is based on the comparison, none of the characters will be displayed. The character "v" beside a figure indicates that the result is significantly better than the baseline classifier (first classifier in the test). Meanwhile, the character "*" indicates a poor result compared to the baseline classifier. However, an indicator is absent if the test cannot say it is either better or worse.

Experiment A examined four different classifiers from Section 1 of the previous experiment (Group 1). The best and worst classifiers were identified, as well as the two most widely used classifiers in literature when modelling the ozone concentration (MLP and SMOreg). The results of the evaluation illustrated that the accuracy of Random Forest is significantly better than that of the other classifiers.

This result supports the conclusion of the Group 1 experiments. On other hand, *Experiment B* evaluated four classifiers, which were, the two most accurate in

each of Sections 1 and 2 of the Group 1 experiments. The results indicate that the performance of bagged Random Forest is significantly better than the rest.

Since some studies, such as [87], have stated that Random Forest as the best performing ensemble classifier, combining Bagging and Random Forest can produce a powerful result, as shown in this experiment.

In *Experiment C*, the evaluation was focused on ensemble learning and the most frequently used single based learning algorithms in predicting ozone concentrations, which are Multiyear Perceptron and Support Vector Machine (SMOreg). The results, as illustrated in Table 5.3, show that MLP and SMOreg have significantly worst prediction result when compared with ensemble techniques. Furthermore, bagged Random Forest has significantly better prediction accuracy.

Table 5.3: Result of Evaluation Experiments

Experiment	Description	Classifier	Result
Experiment A	Evaluate four different classifiers from Section 1 in Group 1 experiments	Random Forest	0.91
		M5Rule	0.89*
		Decision Stump	0.55*
		Multilayer Perceptron	0.89*
		SVM for regression (SMOreg)	0.84*
Experiment B	Evaluate the best two classifiers resulting from the Group 1 experiments	Random Forest	0.91
		Random Subspace	0.90
		Bagging with Random Forest	0.92v
		Bagging with Random Subspace	0.91
Experiment C	Evaluate the most used classifiers in predicting ozone construction with three different ensemble methods	Random Forest	0.91
		Bagging with Random Forest	0.92v
		Bagging with Multilayer Perceptron	0.90
		Random Subspace	0.90
		Multilayer Perceptron	0.89*
		SVM for regression (SMOreg)	0.84*

5.6.2 Group 2

This group of experiments was designed to investigate the performance of heterogeneous ensemble learning algorithms in the modelling of O_3 concentrations. The results of the Group 1 experiments revealed that homogeneous ensemble learning approaches outperformed the single base learners. Specifically the Group 2 experiments use heterogeneous ensemble learning algorithms, Voting and Stacking. Since there is no rule in specifying how many base classifiers should be considered when applying both approaches [93], the Group 2 experiments test the use of combinations of three, four, five and six single base learning algorithms (illustration of Group 2 criteria can be found Section 5.5) .

Table 5.4: Results Of Group 2 Experiments

Based Classifiers		Voting				Stacking				
		CC	MAE	RMSE	RAE	Meta Classifier	CC	MAE	RMSE	RAE
1	**Random Forest, M5Rules, RandomSubspace**	0.91	7.33	10.14	39.16%	**Random Forest**	0.89	7.92	10.90	42.27%
						M5P	0.91	7.15	9.85	38.19 %
						Bagging	0.91	7.36	10.19	39.31 %
						Linear Regression	0.91	7.15	9.84	38.16 %
						Decision Table	0.90	7.65	10.48	40.85 %
2	**Random Forest, M5Rules, MLP,**	0.91	7.56	10.31	40.39%	**Random Forest**	0.90	7.85	10.72	41.89 %
						Linear Regression	0.91	7.14	9.80	38.11 %
						Decision Table	0.90	7.62	10.42	40.69 %
						RandomTree	0.84	10.20	13.80	54.49 %
3	**Random Forest, M5P, M5Rule, KStar, Liner Regression,**	0.91	7.23	10.09	38.62%	**Random Forest**	0.91	7.42	10.20	39.65 %
						Linear Regression	0.92	6.83	9.51	36.49 %
						Decision Table	0.91	7.51	10.28	40.08 %
						REP Tree	0.90	7.51	10.41	40.12 %
						MLP	0.90	8.09	10.74	43.21 %
4	**Random Forest, M5P, Random Tree, M5Rules, MLP,**	0.91	7.38	10.12	39.38%	**Random Forest**	0.90	7.65	10.48	40.83 %
						Linear Regression	0.92	7.07	9.73	37.75 %
						Decision Table	0.91	7.53	10.29	40.18 %
						Random Tree	0.84	10.09	13.89	53.89 %
5	**Decision Table, REPTree, KStar, Liner Regression,**	0.90	7.73	10.72	41.27%	**Random Forest**	0.90	7.65	10.74	40.86 %
						Linear Regression	0.90	7.35	10.37	39.26 %
						LWL	0.78	11.79	15.25	62.94 %
						MLP	0.87	9.14	12.12	48.81 %
6	**Random Forest M5Rule RandomSubspace MLP KStar**	0.91	7.38	10.23	39.42%	**Linear Regression**	0.92	6.82	9.48	36.42 %
						MLP	0.90	8.25	10.88	44.06 %
						REPTree	0.91	7.44	10.23	39.72 %
						IBK	0.85	9.73	13.33	51.97 %
7	**Decision stump, simple linear regression, LWL, Decision Table**	0.79	12.06	16.03	64.38%	**Random Forest**	0.86	9.18	12.49	48.99 %
						Linear Regression	0.81	10.44	14.11	55.75 %
						Kstar	0.85	9.54	12.99	50.92 %
						M5Rule	0.85	9.37	12.63	50.04 %
8	**Random Forest, M5Rule, Decision Stump simple linear regression, LWL, MLP,**	0.88	9.42	12.73	50.32%	**Random Forest**	0.90	7.64	10.50	40.81 %
						Linear Regression	0.91	7.13	9.79	38.09 %
						SMOreg	0.91	7.12	9.80	38.02 %
						IBK	0.84	10.03	13.68	53.56 %
						M5P	0.91	7.13	9.79	38.06 %

When comparing the performance accuracy obtained using voting with different ensembles of single base learning algorithms it was revealed by experiments 1-6 that the use of learning algorithms that performed amongst the

best when used as single learning algorithms in Group 1 experiments, gave a clearly enhanced accuracy as compared with when a mixture of best and worst single learners was used in the voting ensemble (Experiments 7 and 8). As long as the best single learners are utilised in the learning ensemble, there is no conclusive evidence to prove that either an increase of the number of classifiers (Experiments 1, 2 against 3,4), use of classifiers from the same group (e.g. only tree base classifiers in Experiments 2 and 4) or a mixed group of single learning algorithms (Experiments 5 & 6 of Group 2) will result in an overall performance enhancement. However, when the ensemble includes a mixture of good and poor classifiers, the higher the number of algorithms in the ensemble, better would be the overall performance (compare Experiment 7 and 8).

In the Group 2 experiments, Stacking used the same base learning algorithms at Layer 0 used in Voting (see Table 5.4). As, the meta classifier (i.e. Layer 1 learning algorithms) Stacking uses a further single base classifier from those experimented in Group 1. When comparing the overall accuracy figures obtained for Stacking as tabulated in Table 5.4, it is revealed that the use of Linear Regression as the meta classifier gives the best ultimate accuracy figures for Stacking. It clearly improves the accuracy obtained by the ensemble of classifiers used at Layer 0. This observation is in line with the observations made in studies of [94] and [93]. [93] showed that as most of the learning is completed in Layer 0, a simple learning algorithm such as Linear Regression will perform best in finally concatenating the learning experience of Layer 0 within Layer 1.

One further interesting observation is revealed when comparing the results of Experiments 7 and 8. Comparing the learning algorithms used in the Layer 0 ensemble of the two experiments it is seen that Experiment 7 does not involve the best single learning algorithm (from Group 1 results), Random Forest. Further Experiment 8 contains two additional learning algorithms as compared to Experiment 7. Although, in both cases, the use of Linear Regression as the meta classifier has improved the overall accuracy, in Experiment 7, when Random Forest is used as the meta classifier, it outperforms the Linear Regression model. Given that Random Forest was not a part of the Layer 0 ensemble, this proves its impact when it is then used as the meta classifier. In fact, in Experiment 7, Linear Regression performs worst as compared to all other meta classifiers

experimented against. This concludes that the use of a simple learning algorithm such as Linear Regression as the meta classifier in Stacking is only justified if the Layer 0 ensemble contains a collection of best single base learning algorithms.

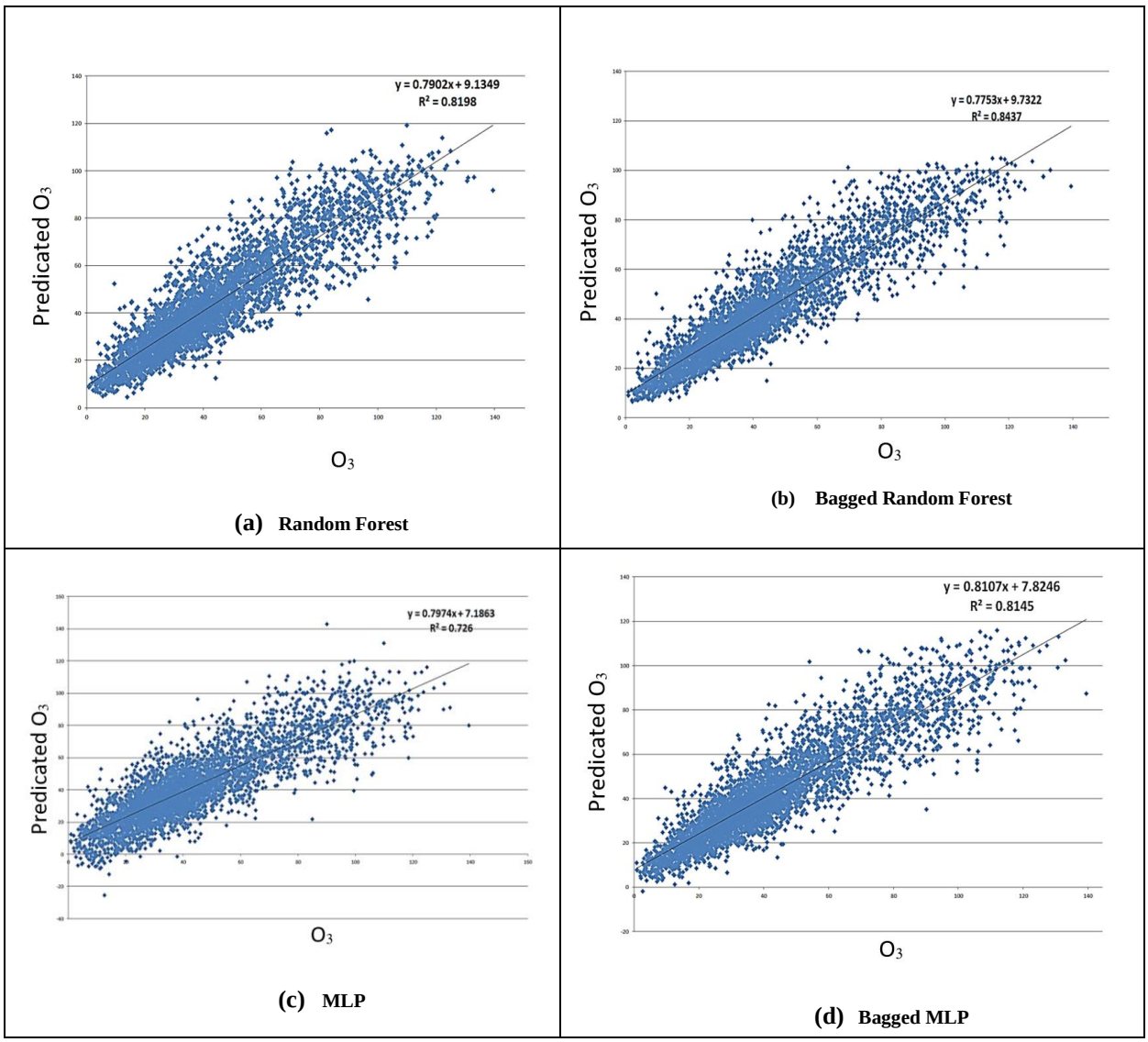

Figure 5.2 (Part 1): Prediction Scatter Graphs

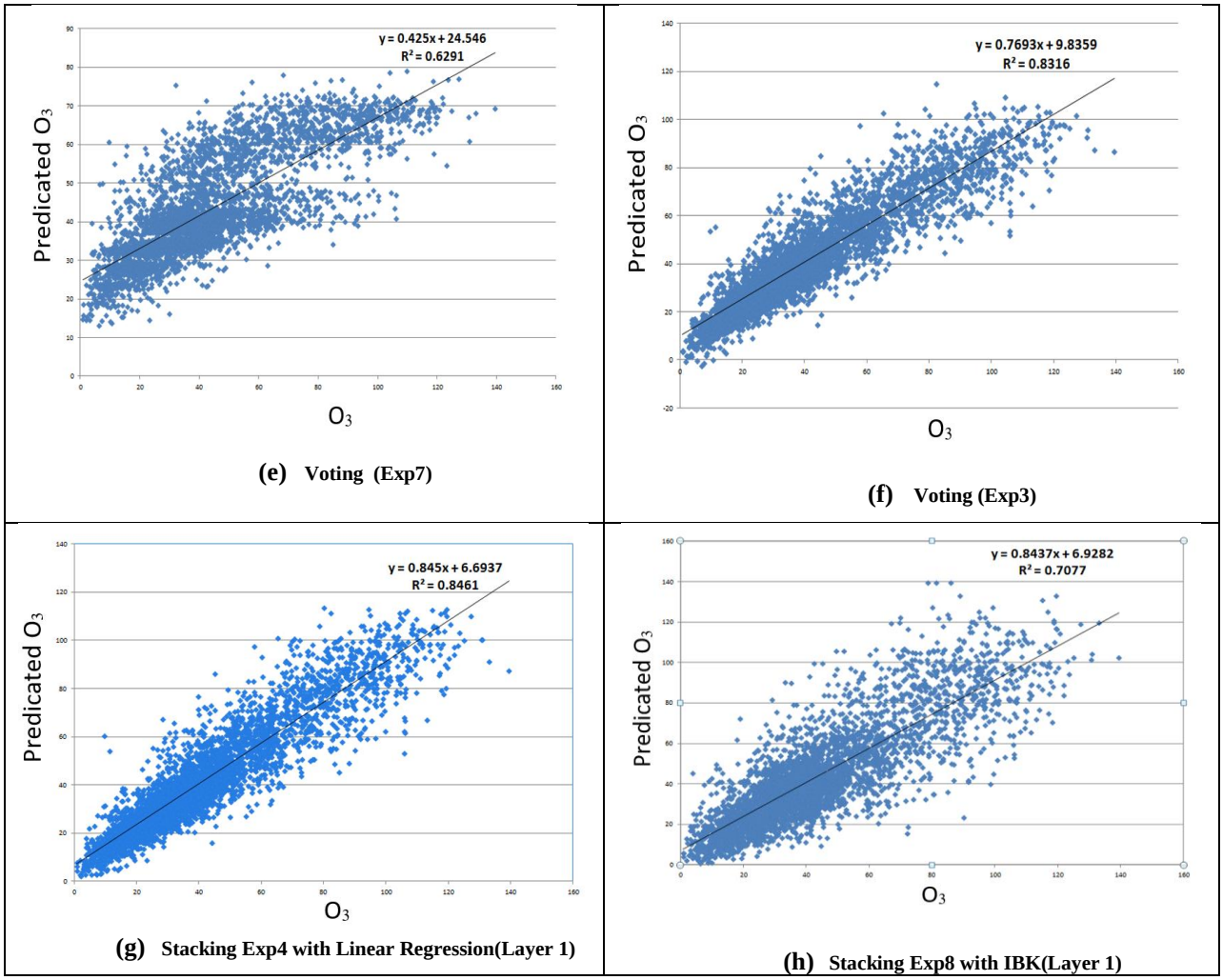

Figure 5.2 (Part 2): Prediction Scatter Graphs

Figure 5.2 illustrates the use of prediction scatter graphs to visualize the impact of using the three ensemble learning approaches, Bagging, Voting and Stacking. Comparing results in (a) and (b) (and (c) and (d)) of Figure 5.2 shows the impact of using Bagging to decrease the scatter of points, especially at lower ranges of ozone concentrations where the density is high. Thus Voting increases prediction accuracy. A comparison of (b) against (d) reveals the ability of Random Forest to reduce scatter (i.e. increase prediction accuracy) against MLP.

A comparison of (e) against (f) of Figure 5.2 reveals the impact of using the best single learning algorithms in the learning algorithm ensemble of voting. Figure (e) depicts the results of Experiment 7 that uses a combination of best and worst single classifiers shows a significantly high amount of point scatter.

Comparing the results of (g) against (h) of Figure 5.2 reveals the ability of simple learning algorithms such as Linear Regression to improve overall prediction accuracy in stacking when used as the meta classifier. The figures illustrate that Linear Regression creates less scatter of points as against IBK.

5.7 Conclusion

This chapter has presented an investigation on the use of three meta learning algorithms (Bagging, Voting and Stacking) to predicate ground level ozone. The prediction was based on concentrations of seven gases (NO_2, SO_2, and BTX (benzene, toluene, o-,m-,p-xylene) and six meteorological parameters (ambient temperature, air pressure, wind speed, wind direction, global radiation, and relative humidity). The use of several widely used single base classifiers have been experimented and compared with the use of the three ensemble classifiers. The results have shown significant improvement in the model accuracy when the meta learning ensemble classifiers were used. The highest prediction accuracy in terms of correlation coefficient was obtained when the ensemble learning meta classifier, Bagging, was used with Random Forest and the base classifier and when ensemble classifier Stacking was used with Linear Regression as the Layer 1 classifier.

The work presented in this chapter proposes invaluable, novel and more efficient learning approaches to the air pollution prediction research community, who have traditionally used popular single base learning algorithms such as neural networks and linear regression.

CHAPTER 6

Optimising the use of Bagging in Modelling Ground Level Ozone Concentration

It is noted that all machine learning algorithms consist of a number of algorithmic parameters/settings that can be adjusted to obtain the optimum performance of the learning algorithms utilised. In the experiments conducted in Chapter 5, in order to leave the investigations in their simplest format, a decision was made to always use the default parameters of the tested algorithms as defined by the WEKA toolkit. Further feature selection/reduction, i.e. minimization of the number of attributes that models the concentration of atmospheric ozone, may lead to increased prediction accuracy. Although the above two aspects were not investigated within Chapter 5, it concluded that in general, Ensemble Learning Algorithms, outperformed the more commonly used single learning algorithms, such as the support vector machines and neural networks. In particular the performance of the Ensemble Learning Algorithms, Random Forests, and Bagging were investigated in detail and shown to produce encouraging levels of performance accuracy.

6.1. Research Motivation & Overview

In order to further investigate the optimal use of Ensemble Learning algorithm, Bagging, in the prediction of ground level ozone, this chapter carries out the fine tuning of the said algorithm by the use of model parameter based optimisation techniques. The use of a number of different feature reduction/filtering approaches is investigated in detail. The use of Bagging is investigated in detail using the base learning algorithms, Random Forest (classified as a homogeneous ensemble learning algorithm and proven in Chapter 5 to be one of the most efficient learning algorithms) and the popular single learning algorithms, Support Vector Machines and Artificial Neural Networks. The investigations conducted in this chapter provide conclusive evidence that such optimisations result in further improvement of the basic models investigated and recommended in Chapter 5.

Using a parameter based optimisation strategy similar to that proposed for Bagging in this chapter it is be possible to optimise the performance of the two layer, heterogeneous ensemble learning algorithms, Voting and Stacking, investigated in detail in Chapter 5. However, such studies are not the focus of this chapter.

In addition, research presented in this chapter investigates the impact of using attribute selection/reduction when using all algorithms investigated in detail, namely, Random Forests, Support Vector Machines, Artificial Neural Networks (ANN), bagged Random Forests, bagged Support Vector Machines and bagged ANN.

For clarity of presentation this chapter is divided into several sections. Apart from this section which provided the motivation behind the research to be presented, Section 6.2 presents the methodology of research conducted. Section 6.3 illustrates the modelling procedure of ozone concentration. On the other hand, Section 6.4 provides the experimental results and a detailed analysis of the results. Finally, Section 6.5 concludes with an insight into future research.

6.2. Experimental Methodology

Figure 6.1 illustrates how experiments conducted in Chapter 5 are related to the experiments to be conducted under the research remit of this chapter. A selected set of classifiers tested under default settings of algorithmic parameters in Chapter 5 and using all input attributes for modelling, i.e. the six selected classifiers (MLP,SMOreg, RF, bagged MLP, bagged SMOreg, and bagged RF) are optimised based feature selection/reduction algorithms and parameter based optimisation techniques. Four attribute filters namely, CFS- Best First, CFS – Greedy Stepwise, Relief Attribute Evaluation and Principle Component Analysis (refer to Section 3.7.1 for more detail) implemented within WEKA data mining toolkit are used for feature/attribute selection. Further for optimal parameter selection within each of the tested models, the algorithm ‘CVParameterSelection’ (see Section 3.7.2) as implemented in WEKA has been used.

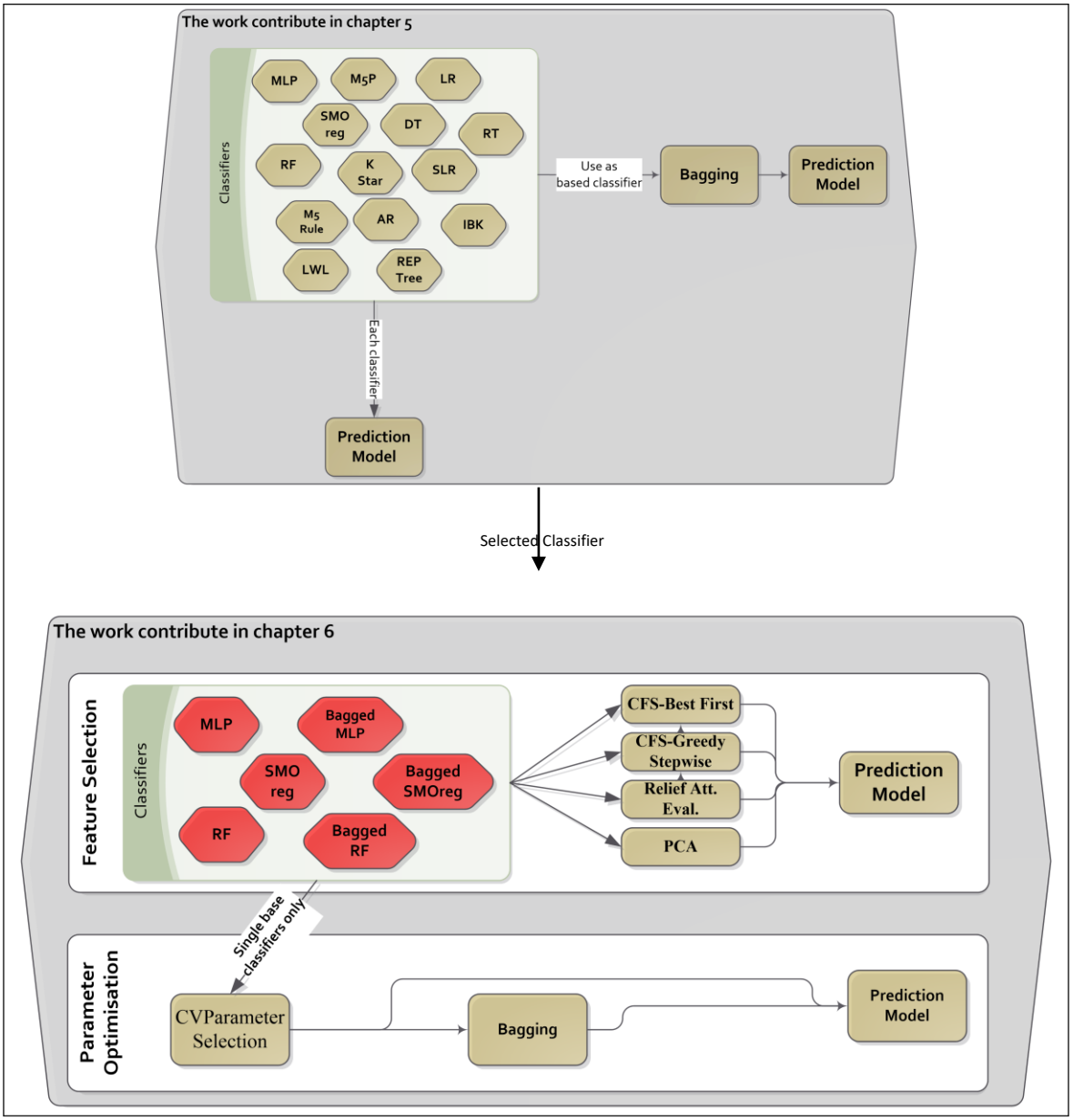

Figure 6.1: Experimental Methodology

6.3. Modelling the Ozone Concentration

Although the performance of a large number of other classifiers and classifier combinations were investigated in a preliminary study, the detailed analysis of the performance of only the six algorithms mentioned above i.e. the SVM, ANN and Random Forest when used with and without Bagging, is presented in this chapter. The accuracy of the algorithms is evaluated using two widely used evaluation metrics, Correlation Coefficient and Mean Absolute Error (see Section 3.6 of Chapter 3).

To present a fair performance comparison between the classifiers, optimal parameters for each classifier are determined using the CVParameterSelection algorithm prior to conducting detailed modelling. The Explorer experimental GUI environment of WEKA was used to construct individual classifier models using their optimal parameters settings. The performance of the six different classifiers are analysed and compared, using the same dataset (see Section 4.1) using the Explorer. Tenfold cross validation was used to minimize the effects of chance in dividing the dataset to test and training sub-sets.

6.4. Experimental Results and Analyses

Experiments were conducted to analyse and compare the performance of the six classifiers: MLP (WEKA's ANN implementation), SMOreg (WEKA's SVM implementation), Random Forest (RF), bagged MLP, bagged SMOreg and bagged RF as stated above. Further detailed experiments were also conducted to determine the potential impact of feature reduction / selection and in the selection of classifier parameters in optimising classifiers, in the overall accuracy obtainable from each of the six evaluated classifiers. It is noted that the original readings recorded for wind direction was a measure in the range 0-360 degrees. In order to compensate for the fact that 0 and 360 degree readings mean the same, this study has combined wind direction (WD) with wind speed (WS) to replace them with two orthogonal components WS*cos(WD) and WS*sin(WD).

It is noted that all of the classifiers investigated (i.e. regardless of whether the classifier is of the single classifier type or the ensemble classifier type) consist of a number of input parameters that may have a vital impact on the accuracy of predictions obtainable. Although WEKA provides default parameter values for each classifier, our preliminary experiments suggested that these values do not result in optimised prediction. Therefore, it was vital to select a set of parameters which provide optimal prediction accuracy. For this purpose the use of WEKA's CVParameterSelection filter has been made. Table 6.1 tabulates the prediction accuracy obtainable via each approach in terms of correlation coefficient. The results indicate that the optimal parameter selection has a positive impact only when use the single classifiers MLP (i.e. ANN) and SMOreg (i.e. SVM). When using ensemble

classifiers Random Forest and Bagging, the optimal parameter selection algorithm has no impact, indicated by the accuracy figures that remain unchanged. It is noted that even though the CVParameterSelection filter changes some parameters in its attempt to optimise the accuracy, no change is indicated in comparison to the accuracy obtainable using default settings. For ease of comparison of results presented in Table 6.1.

Table 6.1: Experiments results for parameter based optimisation of the classifiers

Classifier Name	Default settings	Correlation Coefficient	Optimal Parameters	Correlation Coefficient
Bagged RandomForest	Bagging: bag size percent (P)=100 Number of iteration(I)=10 Seed (S)=1 num-slots =1 Random Forest: NumTree (I)=10 NumFeature (K)=0	0.92	Bagging: bag size percent (P)=100 Number of iteration(I)=10 Seed (S)=1 num-slots =1 Random Forest: NumTree (I)=20 NumFeature (K)=0	0.92
Random Forest	NumTree (I)=10 NumFeature (K)=0	0.91	NumTree (I)=20 NumFeature (K)=0	0.92
Bagged MLP	Bagging: bag size percent (P)=100 Number of iteration(I)=10 Seed (S)=1 num-slots =1 MLP: Learning Rate (L)=0.3 Momentum /(M)=0.2 Hidden layer= a (attribute/class)/2	0.90	Bagging: bag size percent (P)=100 Number of iteration(I)=10 Seed (S)=1 num-slots =1 MLP: Learning Rate(L)=0.1 Momentum (M)=0.1 Hidden layer= 5	0.90
MLP	Learning Rate (L)=0.3 Momentum /(M)=0.2 Hidden layer= a (attribute/class)/2	0.85	Learning Rate(L)=0.1 Momentum (M)=0.1 Hidden layer= 5	0.88
SMOreg	C:1.0 Kernal: polyKernel	0.84	C:1.0 Kernel: NormalizedPolyKernel	0.89
Bagged SMOreg	Bagging: bag size percent (P)=100 Number of iteration(I)=10 Seed (S)=1 num-slots =1 SMOreg: C:1.0 Kernal: polyKernel	0.84	Bagging: bag size percent (P)=100 Number of iteration(I)=10 Seed (S)=1 num-slots =1 SMOreg: C:1.0 Kernal: NormalizedPolyKernel	0.89

Table 6.2 summarises overall prediction accuracies obtainable by each classifier presented in terms of the Correlation Coefficient and Mean Absolute Error with both using the default parameter settings of WEKA and with optimised parameter settings.

Table 6.2: Summary of Results –Parameter Based Optimisation

	Default Parameter		**Optimising Parameter**	
Classifier	**Correlation Coefficient**	**Mean Absolute Error**	**Correlation Coefficient**	**Mean Absolute Error**
Bagged Random Forest	0.92	7.08	0.92	7.05
Random Forest	0.91	7.52	0.92	7.16
Bagged MLP	0.90	7.64	0.91	7.27
MLP	0.85	9.81	0.88	8.51
SMOreg	0.84	9.54	0.89	8.05
Bagged SMOreg	0.84	9.54	0.89	8.04

Figure 6.2 illustrates graphs representing the actual ozone concentration versus the predicted ozone concentrations. The graphs illustrate the better prediction capability of Bagged Random Forest classification approach as compared to the others. Data points lie closer to the line of approximation (less spread) than in the other graphs indicating a better overall prediction accuracy.

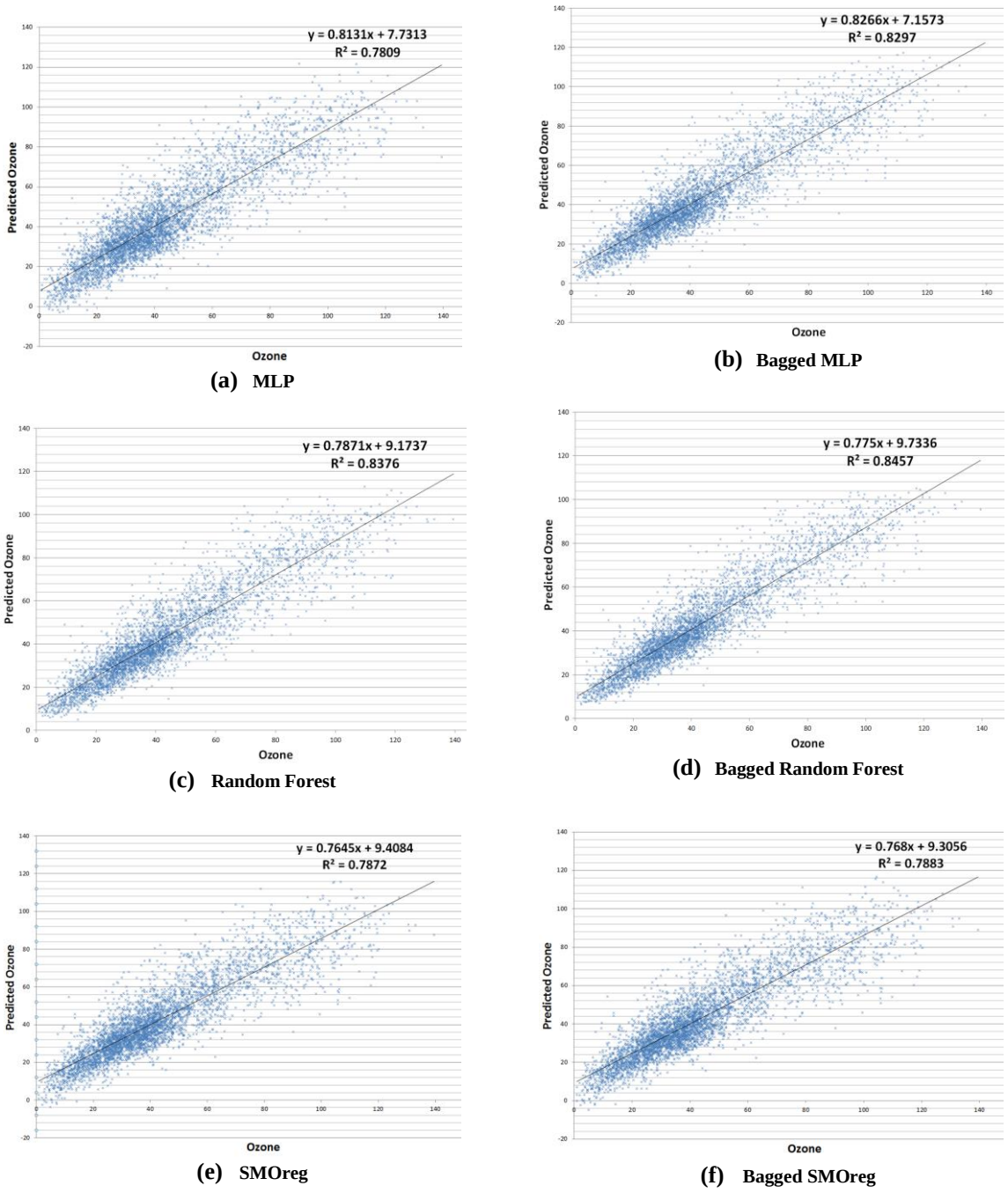

Figure 6.2: Scatter Plots of the actual and predicted Ozone for 6 Models

Table 6.3 tabulates the accuracy values obtained when using four different attribute filtering approaches implemented within WEKA, namely, CFS Subset Evaluator, with Best First and Greedy Stepwise search, Relief Attribute Evaluator and Principle Component Analysis (for detail refer to Chapter 3). The results indicate that no improvement of accuracy is achieved in comparison with using all attributes. This work also investigated the impact of removing wind

direction from being considered, taking only the wind speed into account (from the original data recorded). It was seen that the wind direction has negligible impact on the ozone concentration prediction accuracy. This is justifiable as the measurements for ozone was done across the road, i.e. at its source, as it was vehicular traffic that was suspected to create the ozone from the nitrogen dioxide emissions from the vehicles.

Table 6.3: Results of applying feature/attribute selection

	MLP	SMOreg	Random Forest	Bagged MLP	Bagged SMOReg	Bagged RandomForest
CFS-Best First	0.82 (-3)	0.82 (-2)	0.89 (-3)	087 (-3)	0.82 (-2)	0.90 (-2)
CFS-Greedy Stepwise	0.81 (-4)	0.82 (-2)	0.88 (-4)	0.86 (-4)	0.82 (-2)	0.90 (-2)
Relief Att. Eval.	0.83 (-2)	0.83 (-1)	0.91 (-1)	0.89 (-1)	0.83 (-1)	0.92 (0)
PCA	0.84 (-1)	0.83 (-1)	0.87 (-5)	0.89 (-1)	0.83 (-1)	0.89 (-3)
Using All Attributes	**0.85**	**0.84**	**0.92**	**0.90**	**0.84**	**0.92**

Due to the prediction algorithm adopted by Bagging (see Section 3.4.2) it resolves the data over-fitting problem associated with most classifiers, in this case with MLP and SVM in particular. This is the reason for the significantly better prediction accuracies obtainable from using the Ensemble Classifier Bagging as against the accuracies obtainable from the traditional single classifiers commonly used in predicting ozone, ANN and SVM. In addition, there is no substantial improvement obtainable when using feature selection for Random Forest. This is due to fact that RF algorithm (see section 3.4.2) uses a approach of feature selection when it builds the model. This inherent feature selection negates the need of any feature selection outside the algorithm's operation.

6.5. Conclusion

The chapter has compared the performance of six selected machine learning algorithms in predicting the ground level atmospheric ozone concentrations. The prediction was based on concentrations of seven gases (NO_2, SO_2, and BTX

(benzene, toluene, o-,m-,p-xylene) and six meteorological parameters (ambient temperature, air pressure, wind speed, wind direction, global radiation, and relative humidity). Results prove the ability of ensemble learning algorithms, Random Forests and Bagging to perform significantly better than the most widely used learning algorithms in literature for the prediction of ozone concentrations, namely the Artificial Neural Networks and Support Vector Machines. In addition, the results show bagged with Random Forest gives the best performance for the dataset for which the investigation was carried out and the parameters were adopted are listed in Table 6.1. Specifically, the research presented in this chapter used parameter based optimisation techniques for the optimum parameter selection for each algorithm experimented and investigated the possible use of attribute/feature reduction techniques that were both expected to improve prediction accuracy. However, the experimental results and the detailed analysis revealed that only marginal improvements can be gained by adopting the above techniques.

CHAPTER 7

Application of Time Series Analysis in Forecasting Ground Level Ozone Concentration

7.1 Introduction

In contrast to the investigations carried out in Chapters 5 and 6, this chapter investigates the time dependent analysis of variations and trends of ozone concentration, commonly named in literature as 'time series analysis'. For practical relevance of the results produce the focus is to predict 24 hours (i.e. a day) ahead. Two types of time series analysis are conducted, namely, univariate time series analysis and multi-variate time series analysis. Univariate time series analysis refers to forecasting future ozone concentrations based only on known, measured, previous ozone concentrations whereas the multivariate forecasting refers to the prediction of future ozone concentrations based both on the past, measured concentrations of ozone, concentrations of other gases and meteorological parameters that are known to have an impact on ozone formation. As the latter approach is likely to provide more accurate predictions a comparison of the performance of both algorithms has been presented.

Forecasting future concentrations of ozone either based on univariate or multivariate analysis requires the use of historical data of significant time-duration so as to different long term and short term trends and variations can be accurately captured and used in the predictions. Therefore, for this purpose the Sohar University environmental dataset (see Chapter 4) used in Chapters 5 and 6 was deemed to be unsuitable due to the fact that the data was gathered during a relatively short period of time and had missing values that resulted in a significant fraction of data being removed from being considered. Thus the DEFRA database (see Chapter 4) is used for this research and analysis.

For the purpose of experimentation and analysis the Time Series Forecasting (TSF) Toolkit of WEKA [95] was used, that provides access to software implementations that have the flexibility to be adjusted and changed according to analysis preferences. Instead of carrying out a comprehensive analysis using a wide range of machine

learning algorithms, as in the research conducted in Chapter 5, a limited number of candidate machine learning algorithms that were proven to perform optimally in spatial forecasting of ground level ozone in Chapter 6, are investigated in this chapter. The performance of ensemble learning algorithms (Bagging and Random Forest (RF)) has been examined and compared with popular approaches used in literature for time series analysis, i.e., MLP and SVM for regression (SMOreg). It is noted that research conducted to date and presented in literature has not investigated the use of ensemble learning approaches in time-series analysis. Therefore, the research conducted within the scope of this chapter has a relevance and significance and contributes positively to the state-of-art.

For clarity of presentation this chapter is organized as follows: Section 7.2 presents the general concepts/terminology related to time series data analysis and modelling, followed by the presentation of methodology research adopted in this chapter in Section 7.3. Sections 7.4 – 7.6 present the experimental results and a comprehensive analysis of the results. Finally, Section 7.7 summarises and concludes the research conducted.

7.2 Time Series Analysis

Time series analysis and forecasting models have been widely used in research in many application areas during the last decade. The main aim of time series analysis is to study the historical behaviour of a collected dataset in order to use such behaviour to develop a forecasting model which can provide the means for obtaining information about the future (i.e. understanding the past can help to model the future), in advance. Such advance notice for example of ground level ozone concentrations can help in providing general public with timely warnings for excess or dangerous levels of ozone concentrations.

7.2.1 Time Series Data

A time series dataset is defined as a set of data measured/captured periodically (e.g., hourly, daily, monthly, or yearly) and can be continuous or discrete in value. Hence, the data is time dependence and the ordering of data elements is important. In addition, a time series dataset could be either univariate, which contain records of a single variable; or multivariate where the record contains more than one variable.

In general time series data can demonstrate Trends and Cyclical, Seasonal variations [96]. These terms can be defined as follows.

Trend: time series data either increase or decrease over a period of time.

Cyclical: time series data can fluctuate over a long period of time, typically in excess of one year. The data will exhibit an increase and decrease in value during the period.

Seasonal: The structure of the data can repeat itself in similar patterns over a fix periodic time, e.g. monthly, or yearly.

Figure 7.1 illustrate a plot of ozone concentration over two, five month periods, in 2010 and 2015, respectively. The data for the plots have been obtained from the DEFRA dataset. The plot indicates the presence of seasonal and cyclical changes (see Figure 7.2), but not trends.

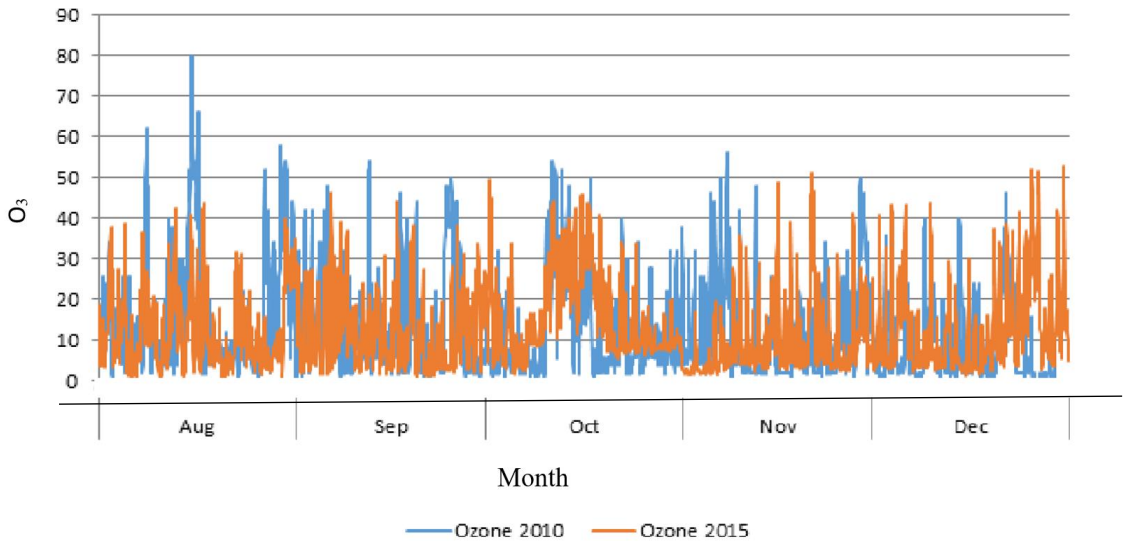

Figure 7.1: Ozone concentration variations for year 2010 and 2015

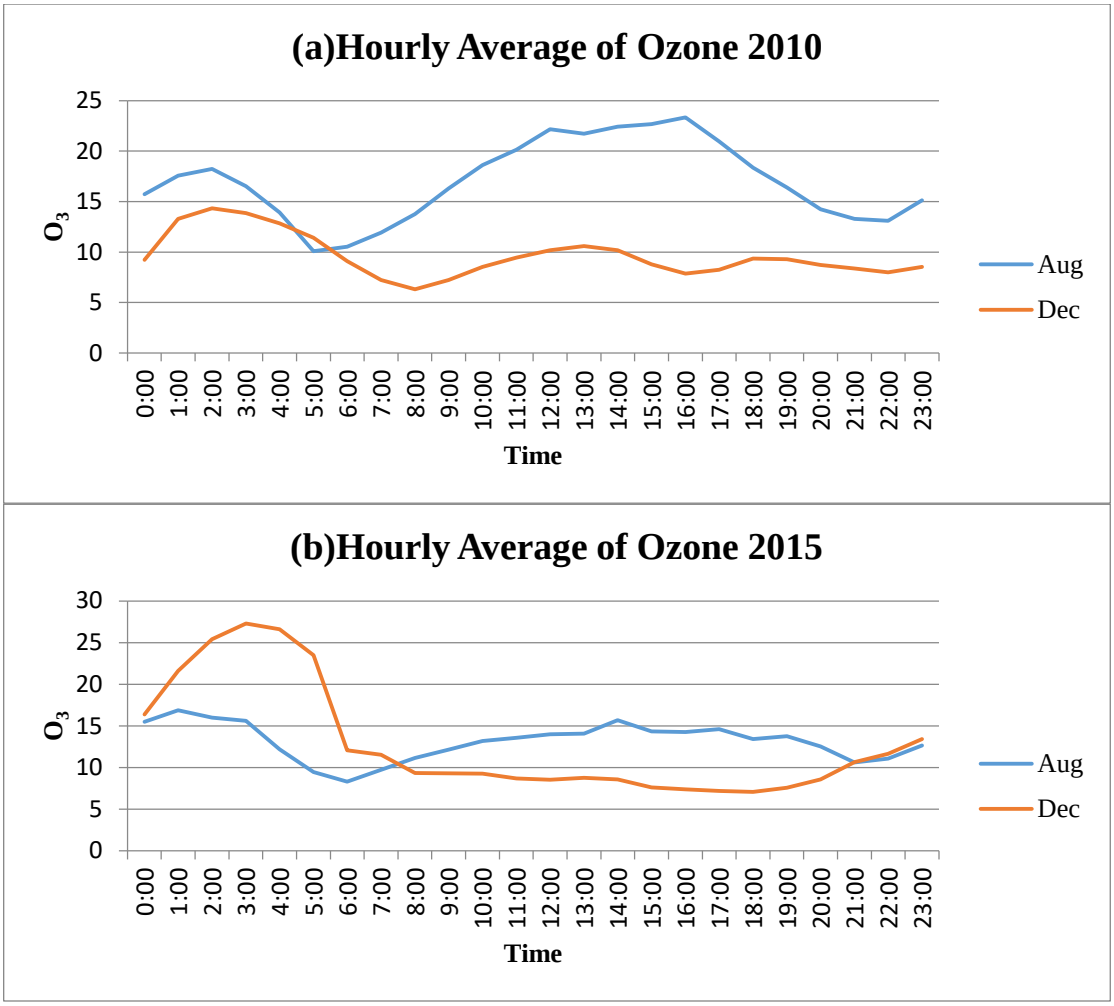

Figure 7.2: Hourly average Ozone concentrations in months, August (summer) and December (winter) in 2010 and 2015)

A dataset used in time series modeling is different as compared to the datasets used in spatial data modelling / data mining due to the fact that a time series dataset has "natural temporal ordering". Thereforein order to make use of the machine learning algorithms implemented within WEKA for forecasting, the input data has to be transformed into a non-time dependent format [97],[98]. Several transformation processers have been adopted to transform a time-dependent dataset into a non-time-dependent dataset. In this work the use of the concept of time lagged variables in order to achieve time independence is made. This is the concept adopted by WEKA's TSF toolkit.

7.2.2 Lagged Variables

In [99] the authors have summarised and presented different methods for pre-processing a time series dataset. One such popular method is by introducing lagged variables to the dataset. Since there are no general rules that specify as to how lagged variable are obtained [34], different approaches to lagged variable preparation has been used in different application domains and by different researchers. In the proposed work, the approach used by WEKA for lagged variable preparation has adopted .

The process of lagged variable preparation can be explained as follows:

A lagged variable is obtained by shifting the target variables X_t by N steps on time space. Where lag N indicate that X_t is holding information of X_{t-n} (see Figure 7.3).

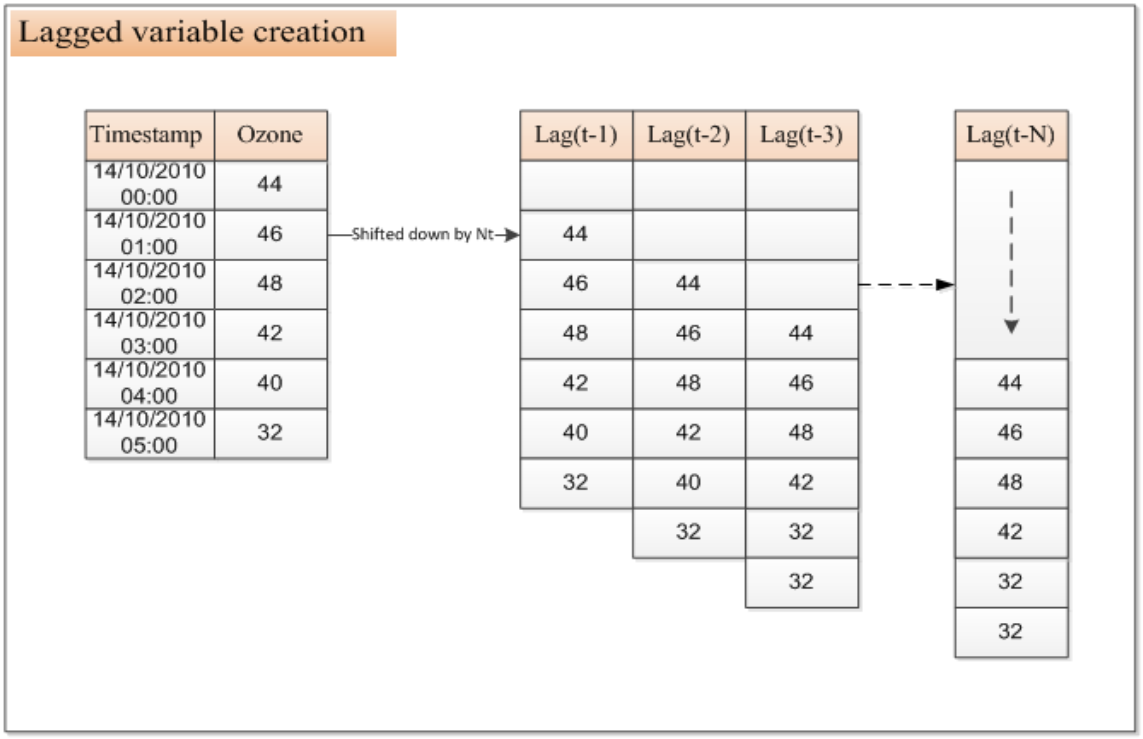

Figure 7.3: Lagged variable creation

Figure 7.4 plots the autocorrelation coefficient [100] of a sample of data depicting the variation of ozone concentration with respect to time at different lags ranging between 1 hour to 24 hours. The graph clearly shows that autocorrelation coefficient drops to approximately 0.5 at a lag of 13hours and

increases above 0.5 at approximately 21 hours. This demonstrates that for most practical cases considering lags from 1-12 hours and just 24 hours will be a sufficient and accurate selection. In the research conducted in this chapter the lags of 1-12 and 24 hours have been used.

Therefor the new fields are added to the original dataset to represent the lag (in this work 13 new fields are included in the dataset). This process will allow the introduction of the historic values at each point in the dataset.

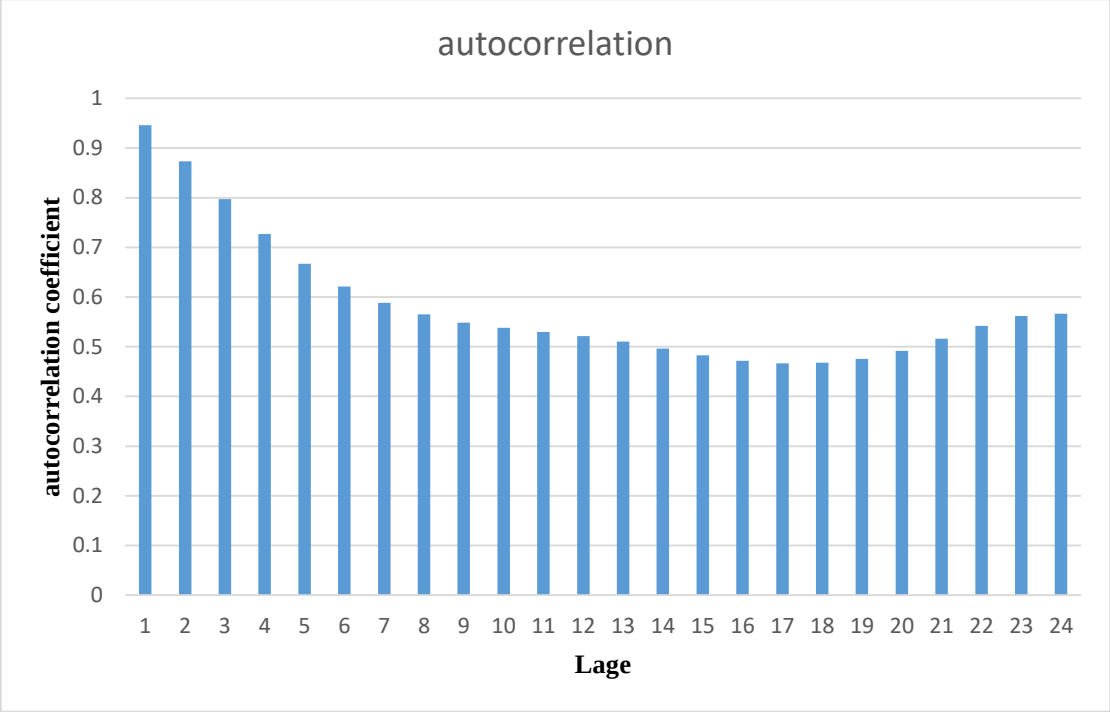

Figure 7.4: Plots of autocorrelation

7.3 Methodology

Within the context of the proposed research several experiments are conducted to model 1-day ahead (i.e. 24 hours ahead at an interval of one hour) forecasting ozone concentration making use of a number of different learning algorithms. The WEKA time series analysis and forecasting toolkit (TSF) is adopted to carry out the forecasting of ozone concentrations. It is noted that WEKA's TSF toolkit is a flexible and powerful tool that can be used for forecasting more efficiently as compared to using other known forecasting models that are based on statistical approaches (e.g. ARMA and ARIMA [100],[101]). In Section 7.2 it was explained how the TSF package automatically handles the temporal ordering of the input data without

recourse to a time-stamp, by introducing inputs as lagged variables to the modelling process. These lagged inputs will be used by several different learning algorithms to model the trends and seasonality of the input dataset and hence to forecast the ozone concentrations 24 hours ahead.

The experimental procedure adopted for time series forecasting making use of the WEKA TSF toolkit is illustrated in Figure 7.5. The forecasting is conducted on the DEFRA dataset (see Section 4.2, Chapter 4) and is fed into the WEKA TSF toolkit (the Forecaster) as the input. In forecasting ozone concentration using the WEKA TSF toolkit, several important parameter selections have to be carried out before the forecaster can start modeling. They are as follows:

(1) Select the target variable as the ozone concentration.
(2) Decide on the number of steps which are required to forecasted as 24, i.e. 24 hrs / 1-day ahead at an interval of one hour.
(3) Select the period as an 'hour'.
(4) Select the level of confidence to be used (95% in proposed experiments).
(5) Select the window size to be used that determines how many past, known ozone concentrations are used for modelling the ozone concentration value to be predicted. (e.g. window size = 24).
(6) Select the learning algorithm to be used in creating the model.
(7) Select the evaluation rule to be adopted.

In multivariate forecasting an additional step has to be performed to overlay the historical meteorological parameter values and concentrations of other gases.

Note from the above parameter selections that when conducting forecasting, the ozone concentration is selected to be the target and Date and the Time has been selected as the time stamp. A total of 24, hourly steps of the ozone concentrations into the future are to be forecasted at 95% confidence. The input data has been lagged by 1-12 hrs and 24 hrs for modelling due to the explanations given with respect to the autocorrelation plots depicted in Figure 7.2 above. The input data has been split as 90% for training and 10% for testing. Note that the evaluation of the results is conducted both within the training set itself and also outside the training set,

i.e. within the test set. The latter approach provides a more practical evaluation scenario.

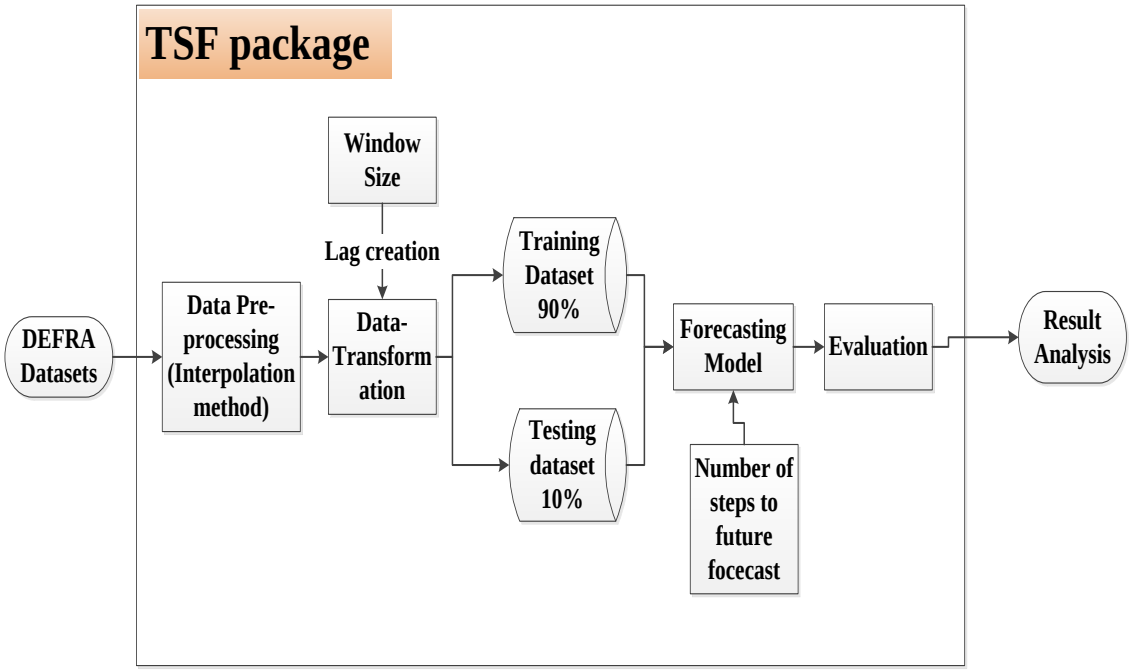

Figure 7.5: The experimental procedure adopted by WEKA TFS toolkit

Several important steps in the forecasting process adopted can be detailed as follows:

7.3.1 Data Pre-processing:

It was observed that the input target data, i.e. the historic and known ozone concentrations had several gaps, i.e. missing values. An interpolation approach (see Section 4.2) implemented within the WEKA TSF package was used as a pre-processing stage for data cleaning.

7.3.2 Data Transformation

As discussed in Section 7.2, the input dataset is transformed to eliminate the temporal ordering by creating time lagged inputs. This transformation is handled within the WEKA TFS toolkit, automatically. Subsequently the time lags to be included in the forecasting process (in the experiments conducted, 1-12 and 24 as explained in Section 7.2) and the window size to be used, should be decided. It is noted that one of the most important factors in the creation of a forecasting model is to select a time window for the model. This window size refers to the

period of training utilised by the forecasting model. Using a narrow window size may not provide sufficient training data for model creation. On other hand, using a wider window size will increase the complexity of the model training process and increase the chances of using irrelevant inputs [102] in training. For the purpose of the proposed research, the window size selected is 24 while the lagged inputs considered in the modelling was, lag1 to lag12 and lag24.

7.3.3 Forecasting Models

Six learning algorithms were selected to build the forecasting models using the WEKA TSF toolkit. These algorithms include MLP, SMOreg, Random Forest, bagged MLP, bagged SMOreg and bagged Random Forest. The models were built using the methodology discussed earlier in this section. The forecasted results are evaluated using two evaluation matrics, the MAE and RAE.

Different experiments have been conducted to find out the learning algorithms that provide the most accurate forecasting of ozone concentrations, 24 hours ahead. Both univariate and multivariate forecasting is conducted. In univariate forecasting, future values of ozone concentrations are predicted based on past ozone concentration data only. However in multivariate analysis the impact of other parameters that are known to have an impact on the formation of ozone such as the concentrations of gases such as SO_2, NO, NO_2, NO_x, and meteorological parameters such as temp, and wind direction/speed are considered in the forecasting of ozone concentrations.

The latest version of the WEKA (version 3.8.0) has been employed to carry out all the experiments; the six selected algorithms were examined using their default parameter settings, with the data set divided as 90% for training and 10% testing.

7.4 Experiments Results and Analysis

As mentioned previously six different machine learning algorithms are used for forecasting. They include MLP, SMOreg, RF, bagged MLP, bagged SMOreg, and bagged RF. All the learning algorithms were operated with their default settings (see

Table 7.1). The experiments were conducted on two datasets; the first set being the entire dataset of ozone concentrations gathered from DEFRA, between 2010 and 2016. Second dataset is a sample of the full DEFRA dataset representing data gathered during a continues period of time (approximately 7 months), without any missing values. Both univariate and multivariate forecasting is carried out for the prediction of ozone concentrations.

Table 7.1: Default Settings for the Classifier Parameters

Classifier	**Default values**
MLP	Momentum= 0.2 LearningRate=0.3 Trainingtime =500 hiddenLayers=(attribs+ classes) / 2 *in this experiment 15 normalizeAttributes= True validationThreshold=20
SMOreg	The complexity parameter C = 1.0 Kernal=PolyKernel filterType= Normalize Training data regOptimizer (The learning algorithm)=RegSMOImproved
RF	bagSizePercent= 100 numIterations =100 maxDepth = 0 (unlimited) numFeatures = 0, which is equal to int(log_2(#predictors) + 1) is used.
Bagging	bagSizePercent= 100 numIterations =10

Two evaluation metrics, the Mean Absolute Error (MAE) and Relative Absolute Error (RAE), have been considered in this work to compare the performance of the six forecasting models. In all experiments conducted, 90% of the dataset is used for training and 10% is used for testing.

7.5 Experiment 1: Univariate Models

In this experiment only ozone concentration data is considered in building the forecasting models. The entire dataset was used and the six different classifiers were

experimented. It is noted that this study have selected the MLP (Multilayer Perceptron is a ANN implementation) and the SMOreg (this is a SVM implementation) as they are the most widely used single classifiers used in literature. Further it is noted that the author has selected the Random Forests (RF) for performance analysis and comparison as it is the simplest form of an ensemble classifier. As the main ensemble learning technique, Bagging (see Section 3.4.2.1) is used. In using Bagging our experiments revealed that the bag size has a significant impact on forecasting accuracy. Therefore, additional experiments (see Section 7.5.1) are conducted to determine the optimal number of bags to be used in the forecasting experiments. Bagging is used with the MLP, SMOreg and RF as the base classifier, providing three additional classifiers for the forecasting tasks.

7.5.1 Number of Bags Optimising for Bagged MLP

The number of bags to be used in the application of Bagging is a factor that needs careful selection. The optimal number of bags to be used depends on the dataset being investigated. Four different experiments were conducted with bag numbers= 3, 5, 10* and 15 (note: * indicates WEKAs default number of bags) to find the best bag size when using bagged MLP. The performance of bagged MLP was investigated when evaluations were done against data from the training set and the test set, respectively. Figure 7.6 plots of MAE vs the 24 hourly (24 step ahead) with graph (a) indicating the results when evaluations were done using data within the training set and graph (b) indicating the results when the evaluations were done using data outside the training set. A careful analysis of the graph (a) illustrates that when evaluations are done using data within the training set both bag sizes 3 and 15 provides the lowest MAE values for all future forecasts. However, graph (b) illustrates that the bag size 3 is not suitable when evaluations are done against the separated, 10% of data. This indicates that the input dataset should be divided into a higher number of bags for different models to be created using the base MLP classifier. The bag number 15 gives the best results.

Within the research context of this thesis the optimal bag number to be used was investigated. When the base classifier is the SMOreg or RF and found that a bag number of 15 remains the optimal number. Therefore for all experiments

conducted a bag number of 15 was used. For a different dataset this value could be different and requires the repetition of the experimental procedure described above to determine a new optimal bag number. Although the investigation carried out above was focused on the creation of a univariate forecasting model, the detailed investigations when using a multivariate forecasting procedure also confirmed the same number of optimal bag numbers.

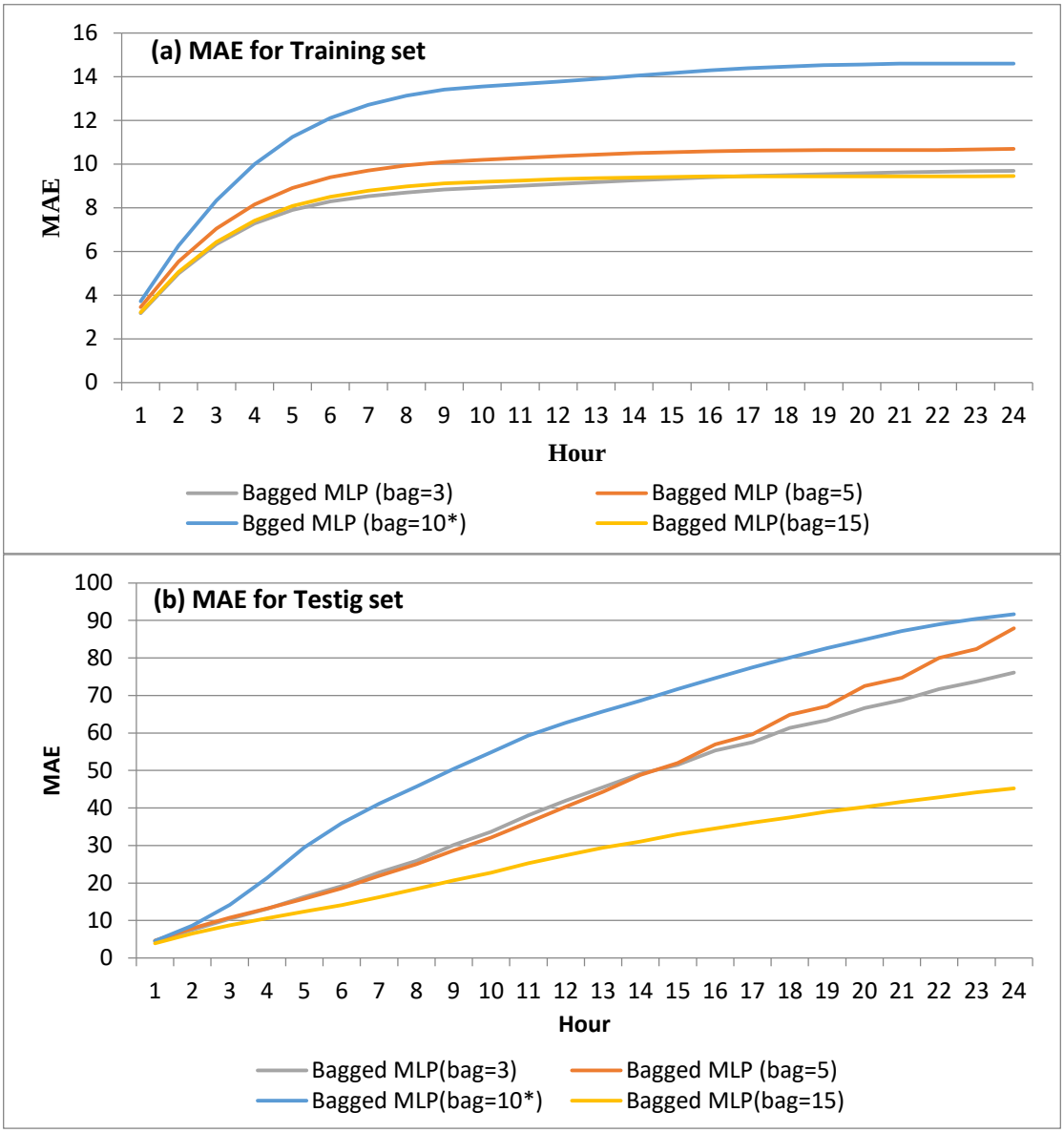

Figure 7.6: MAE of each univariate forecasted hourly ozone concentration for different number of bags for bagged MLP :(a) Result for training set,(b) Result for Test set.

7.5.2 Performance of Univariate Forecasting Models

Figure 7.7 illustrates the forecasting accuracy of all six models experimented, namely the MLP, SMOreg, RF, bagged MLP, bagged SMOreg and bagged RF. Figure 7.7 (a) illustrates the evaluation of results against the training set and Figure 7.7 (b) illustrates the evaluation of results when tested on the separate test set. As expected the accuracy decreases with time, i.e. further the future time for which the ozone concentration is predicted, the error will increase.

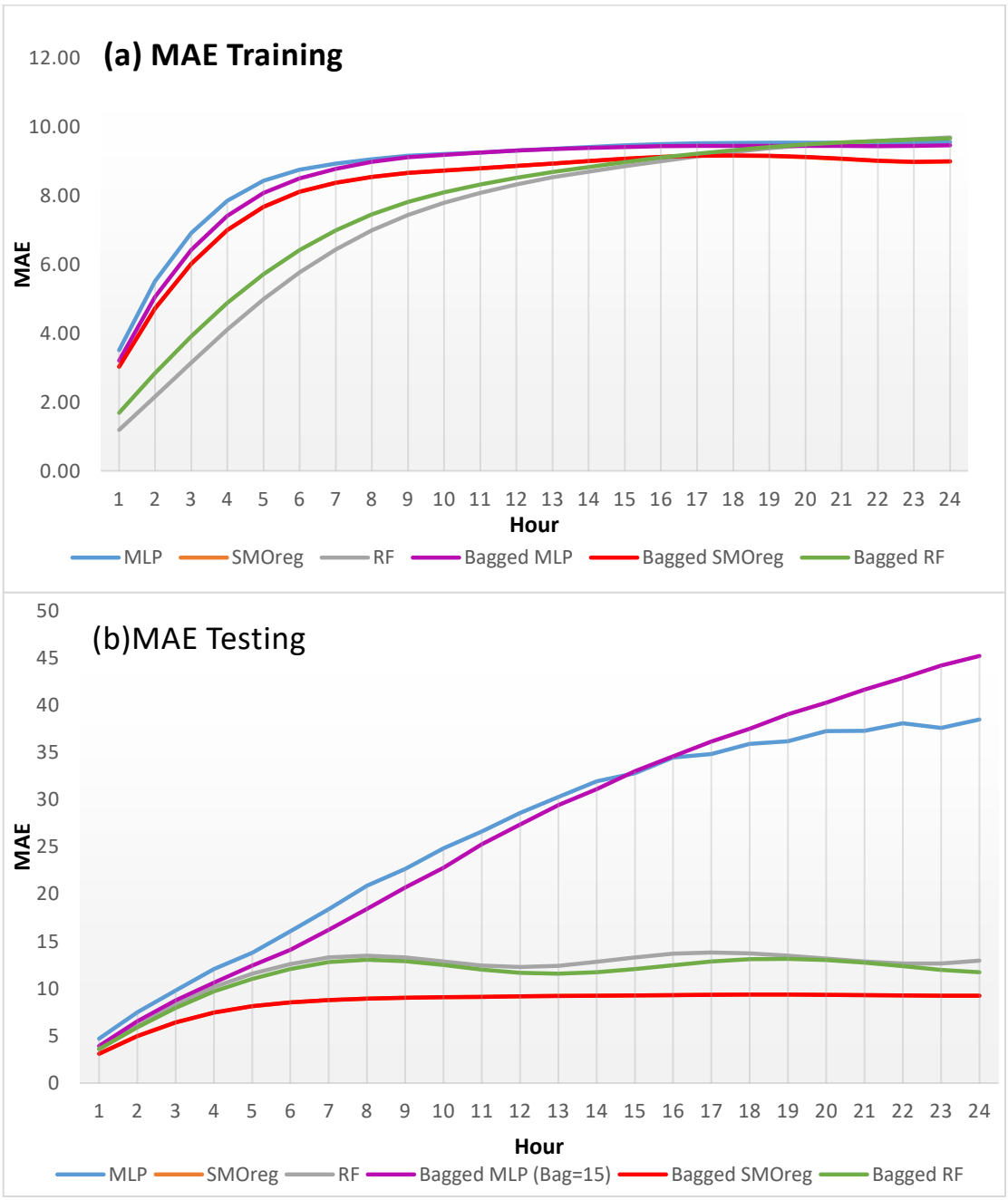

Figure 7.7: Univariate Forecasting – Performance of six classifiers measured in MAE when evaluation is done within the training set (a) and within a separate test set (b).

Note that the MLP, both when used as a single classifier and used as the base classifier of Bagging (bagged MLP), performs worst. Note that in our experiments 15 number of hidden layers was used in MLP and 15 bags were used in Bagging. In the case of testing on separate test data, the error continuously increased with time for both MLP and bagged MLP. This is not the case when the performances of the other four classifiers are considered. It is noted that Bagging marginally improves the results of modelling when used in conjunction with a MLP.

SMOreg and bagged SMOreg performed identically at all times suggesting that Bagging has no impact when used with the SMOreg as the base classifier, suggesting SMOreg's stability as a classifier [91].

The best overall results when testing was done within the training set were obtained with Random Forest (RF), itself an ensemble classifier. Using bagged RF marginally improved the forecasting performance (figure 7.7(b)). The best results when testing on an external test set was obtained by the SMOreg/bagged SMOreg, which were marginally better than the RF/bagged RF.

Figure 7.8 presents the performance accuracy in terms of the Relative Absolute Error. Predictions where the RAE is less than 100% indicates a useful prediction. However if the RAE value is above hundred, it means that rather than forecasting better accuracy results will be obtained by simply taking the prediction to be identical to the ozone concentration the previous day, same time.

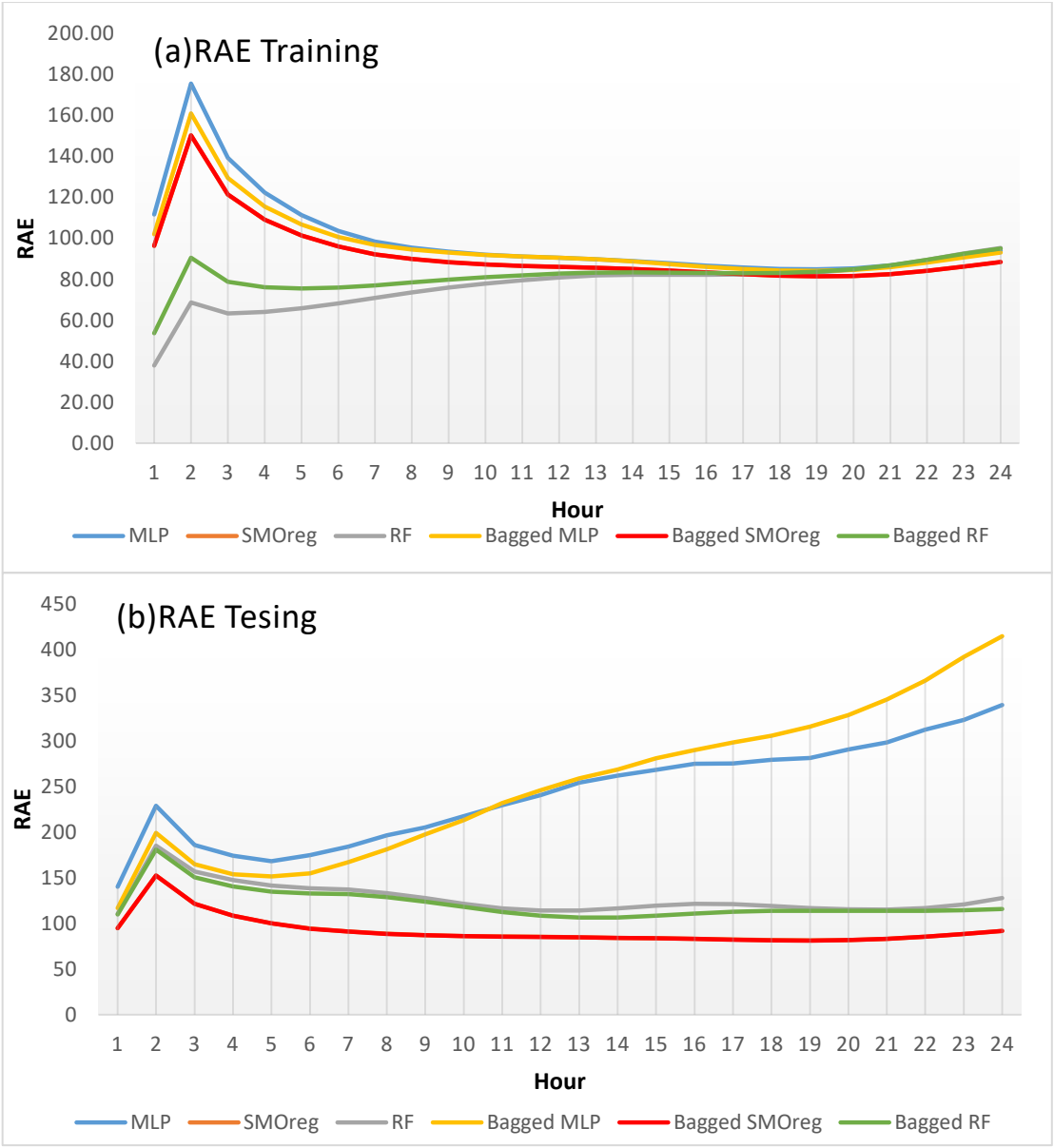

Figure 7.8: Univariate Forecasting – Performance of six classifiers measured in RAE when evaluation is done within the training set (a) and within a separate test set (b).

Table 7.2 includes a screenshot of the actual results displayed by WEKA when bagged MLP was used. The screenshot is provided to provide the reader with an insight into the comprehensive prediction accuracy report WEKA provides.

Table7.2: Screenshot of the bagged MLP Results

	Bagged MLP Training Set							Bagged MLP Testing Set						
	Mean absolute error	Root relative squared error	Direction accuracy	Relative absolute error	Mean absolute percentage	Root mean squared error	Mean squared error	Mean absolute error	Root relative squared error	Direction accuracy	Relative absolute error	Mean absolute percentage	Root mean squared error	Mean squared error
1-step-ahead	3.20	93.30	53.35	102.03	38.99	4.59	21.10	3.93	114.34	59.04	116.84	39.48	6.23	38.85
2-steps-ahead	5.06	140.81	50.15	161.06	69.09	6.93	48.07	6.53	196.17	55.31	199.25	70.78	10.03	100.53
3-steps-ahead	6.42	113.15	47.89	129.54	93.11	8.58	73.53	8.73	190.42	52.92	164.76	97.15	15.09	227.77
4-steps-ahead	7.41	101.73	46.59	115.61	111.12	9.74	94.91	10.62	208.82	50.37	153.72	117.40	21.03	442.08
5-steps-ahead	8.07	94.74	45.81	106.75	122.40	10.53	110.80	12.41	239.16	48.35	151.64	133.87	27.80	772.80
6-steps-ahead	8.50	90.27	45.64	100.68	128.91	11.05	122.06	14.09	272.07	49.02	154.90	145.91	34.46	1187.66
7-steps-ahead	8.78	87.62	45.61	96.76	132.60	11.42	130.31	16.21	326.06	49.80	167.06	159.86	43.56	1897.09
8-steps-ahead	8.97	86.00	45.97	94.55	134.97	11.68	136.37	18.40	378.89	50.05	181.14	174.06	52.36	2741.66
9-steps-ahead	9.11	85.01	46.40	93.07	136.67	11.87	140.85	20.68	435.40	48.56	197.64	187.31	61.55	3788.59
10-steps-ahead	9.18	84.32	46.43	91.86	137.29	12.00	144.00	22.78	481.38	49.37	213.22	199.19	69.32	4805.62
11-steps-ahead	9.24	83.92	46.68	91.10	137.65	12.11	146.55	25.25	537.59	48.35	232.15	216.78	78.78	6205.92
12-steps-ahead	9.30	83.61	46.55	90.50	137.92	12.20	148.87	27.35	576.21	49.68	245.91	231.03	86.17	7425.88
13-steps-ahead	9.35	83.16	46.91	89.78	137.93	12.28	150.90	29.40	613.17	49.56	259.12	244.15	93.57	8755.36
14-steps-ahead	9.38	82.54	46.37	88.72	137.63	12.36	152.74	31.09	634.02	49.21	268.76	258.78	98.72	9746.58
15-steps-ahead	9.41	81.76	45.86	87.45	137.02	12.42	154.31	33.00	658.85	48.77	280.95	274.99	104.40	10898.80
16-steps-ahead	9.43	80.99	46.02	86.18	136.14	12.47	155.59	34.57	672.47	49.02	290.01	288.31	108.25	11718.05
17-steps-ahead	9.43	80.38	46.16	85.08	135.11	12.52	156.67	36.14	690.11	49.35	298.32	301.88	112.49	12654.29
18-steps-ahead	9.44	80.02	46.17	84.30	134.03	12.56	157.69	37.48	702.73	49.26	305.68	313.10	115.48	13336.42
19-steps-ahead	9.44	80.10	46.41	84.11	133.10	12.59	158.61	39.03	721.63	49.51	315.91	326.94	119.20	14207.85
20-steps-ahead	9.44	80.62	46.09	84.59	132.21	12.62	159.28	40.26	739.84	49.39	328.56	330.52	121.44	14748.14
21-steps-ahead	9.44	81.86	46.68	85.95	131.44	12.65	159.97	41.62	770.77	50.57	345.72	337.96	124.34	15460.72
22-steps-ahead	9.43	83.75	47.24	88.14	130.76	12.67	160.58	42.85	806.21	50.35	366.28	346.12	126.74	16062.51
23-steps-ahead	9.44	86.04	47.98	90.70	130.24	12.71	161.49	44.19	851.28	50.62	392.41	351.47	129.55	16782.07
24-steps-ahead	9.45	88.12	48.08	93.02	129.92	12.77	163.03	45.20	883.32	51.46	414.96	357.85	131.16	17202.29

7.6 Experiments 2: Multivariate Model

Univariate forecasting only makes use of the past ozone concentration data for the prediction of future ozone concentration. The section provides experimental results when multivariate forecasting is used for the forecasting of ozone concentrations. Multivariate approach takes into consideration other factors that affects the formation of ozone at a given time, for example temperature, wind speed etc., and concentrations of other gases such as NO_x's that are the gases that decompose to create ozone. In the analysis conducted only used ozone concentration as a lagged variable. This way the future values of external parameters such as wind speed etc., are not taken into account in predicting future values of ozone, but only their past values are considered to ensure that the forecasting is done more accurately. Therefore one would expect multivariate forecasting to produce more accurate results as compared to univariate forecasting.

Two datasets were used in the proposed experiments for conducting multivariate forecasting namely the whole DEFRA dataset that spans over 5 years and a short 7 months (Dec 2012-Auguest 2013) sample dataset, from the full dataset. The two

datasets are used to determine the impact of missing data (regardless of the fact that missing values have been filled or not) on the accuracy of forecasting. Note that whilst the large dataset consists of missing values, the short dataset does not. The short dataset included 5687 instances, which were from approximately 237 days.

7.6.1 Multivariate Forecasting Based on the Short Dataset

Preliminary experiments with multivariate forecasting demonstrated that the optimal bag size to be used for Bagging remains at 15 as an increase beyond 15 only marginally increases accuracy (see Figure 7.9). As in the case of experiments conducted for univariate forecasting 90% of the dataset was used for training and 10% was reserved for testing.

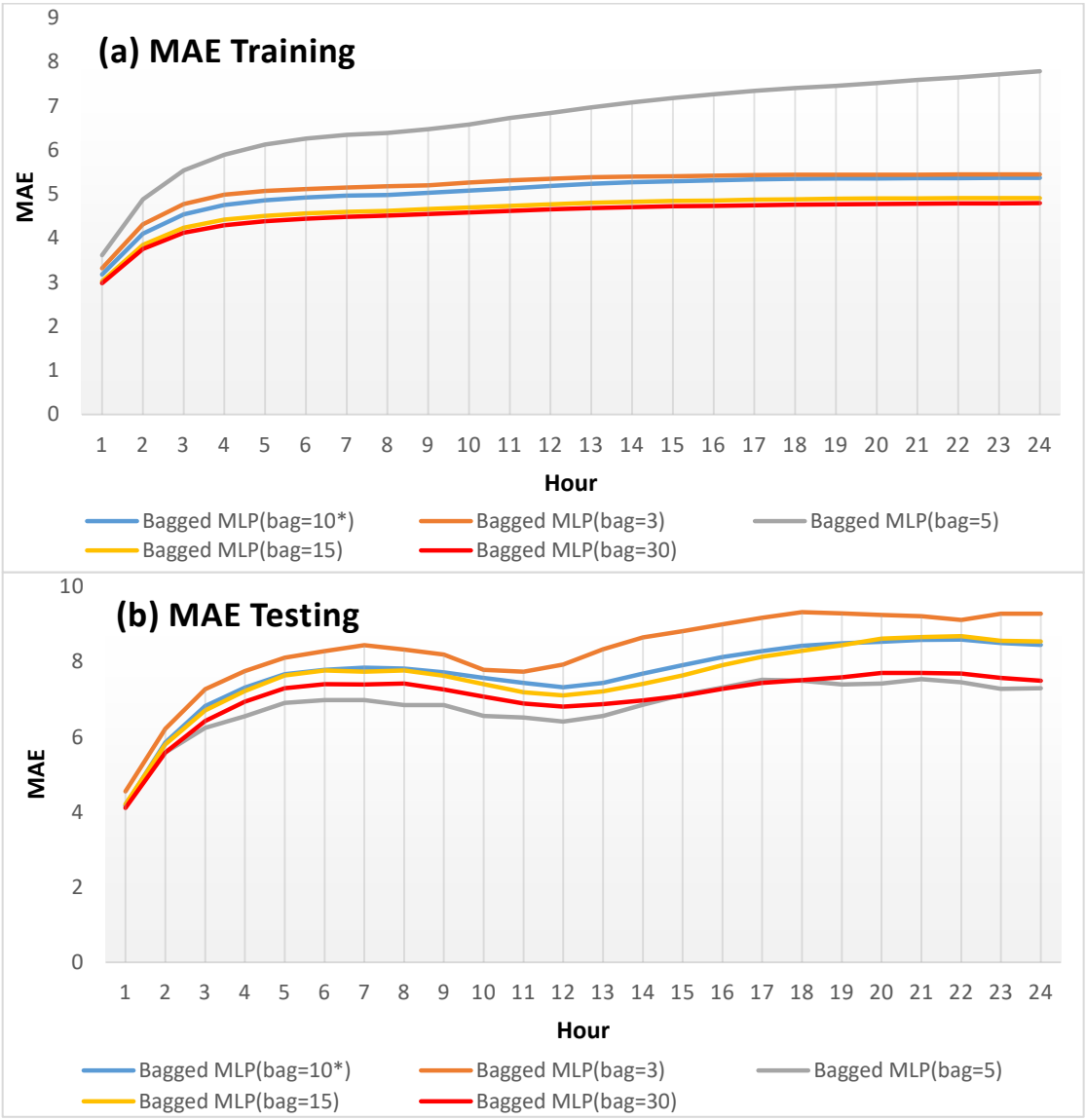

Figure 7.9: MAE of each multivariate forecasted hourly ozone concentration for different number of bags, for bagged MLP :(a) Result for training set,(b) Result for Test set.

Figure 7.10 illustrates the performance of the six multivariate models measured in terms of MAE at hourly intervals, when the evaluation is done within the training set (see Figure 7.10 (a)) and when the evaluation is done on a separate test set (see Figure 7.10 (b)). It shows that different machine learning algorithms perform best under different conditions. When evaluated within the training set, the best performance is given by RF closely followed by bagged RF. When evaluated on a separate test set, the best performance is indicated by bagged MLP closely followed by MLP. RF and bagged RF performs sub optimally when evaluated on a separate test dataset.

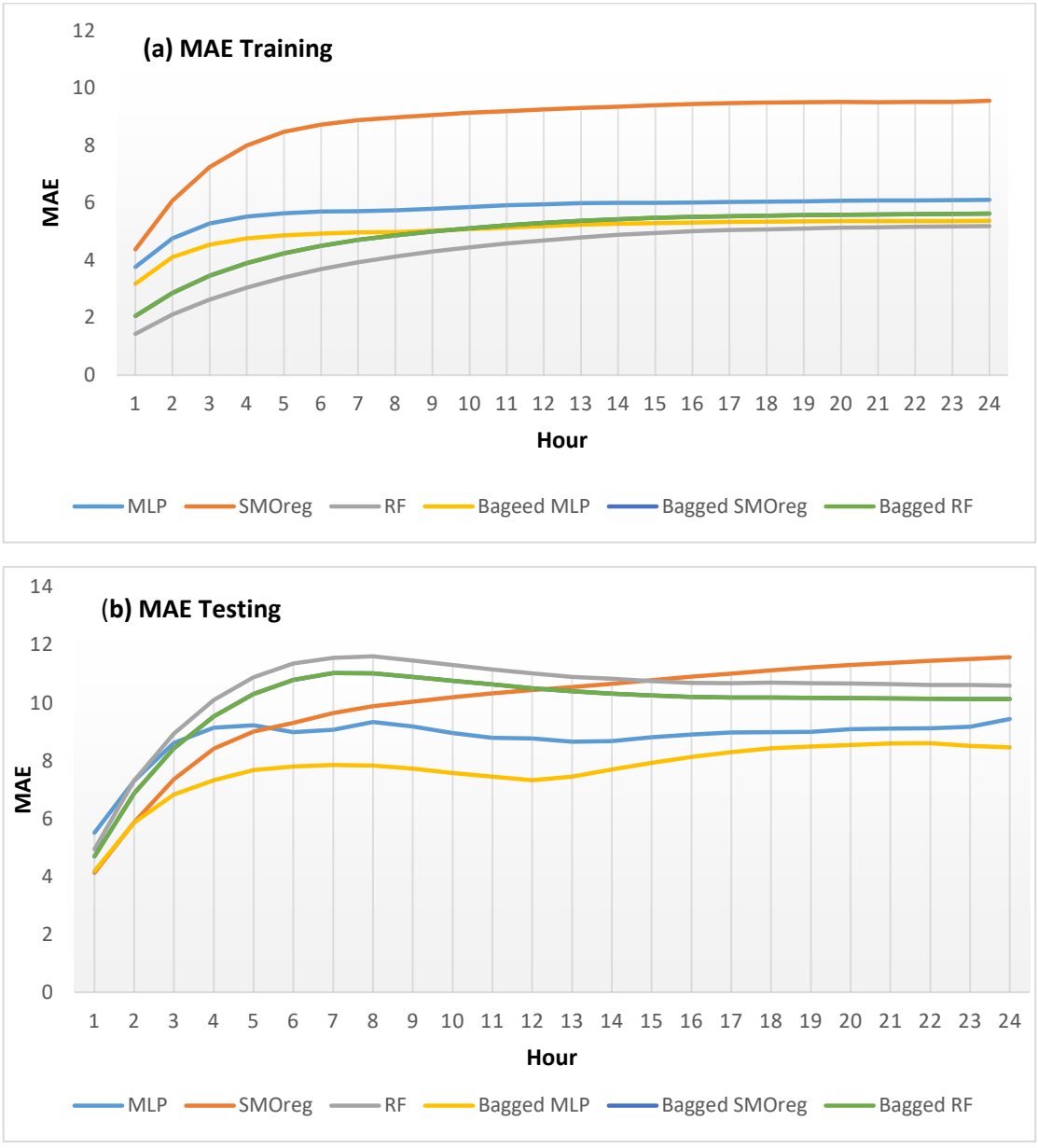

Figure 7.10: Short Dataset: Performance of multivariate forecasting with six different classifiers measured in terms of MAE, (a) when evaluated within the training set and (b) when evaluated within a separate test set

Figure 7.11 illustrates the performance accuracy of the above mentioned multivariate models, measured in terms of the relative absolute error. It is seen that forecasted values have a better accuracy as compared to direct replacement from the previous day's values, for all machine learning algorithms other than when using SMOreg when evaluated within the training test. However when evaluated on a separate test set it is only SMOreg that gives meaningful predictions better than direct replacements, at all hourly intervals within the 24 hour period.

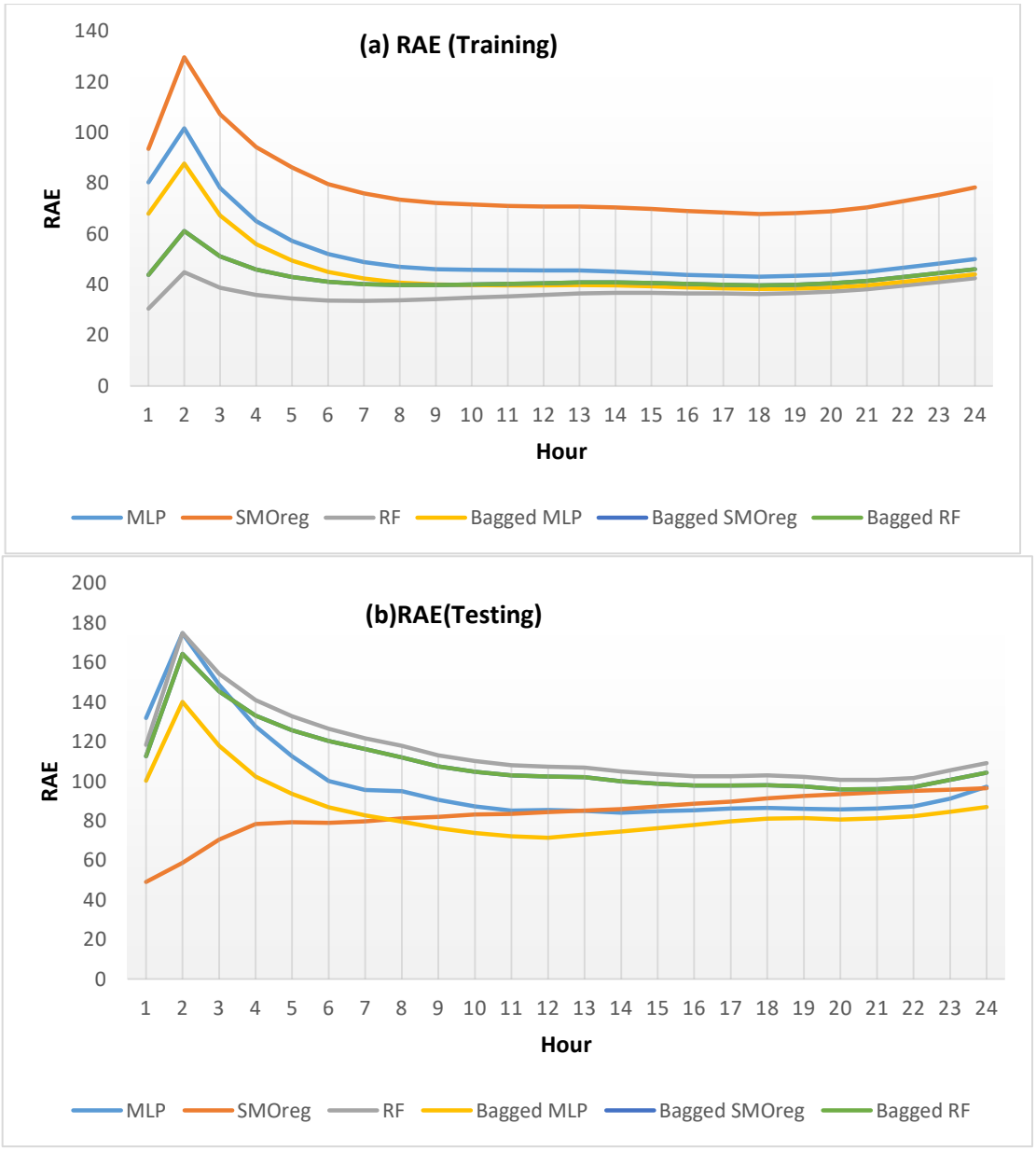

Figure 7.11: Short Dataset: Performance of multivariate forecasting with six different classifiers measured in terms of RAE, (a) when evaluated within the training set and (b) when evaluated within a separate test set.

7.6.2 Multivariate Forecasting Based on the Full Dataset

Figure 7.12 illustrates the performance of the six machine learning algorithms when used in multivariate forecasting, measured in terms of the MAE. Figure 7.12 (a) illustrates the results when evaluated within the training set and 7.12 (b) illustrates the results when evaluated within a separate test set (i.e. within the 10% of the original dataset set aside for independent testing). Both figures indicate that RF and bagged RF performs best.

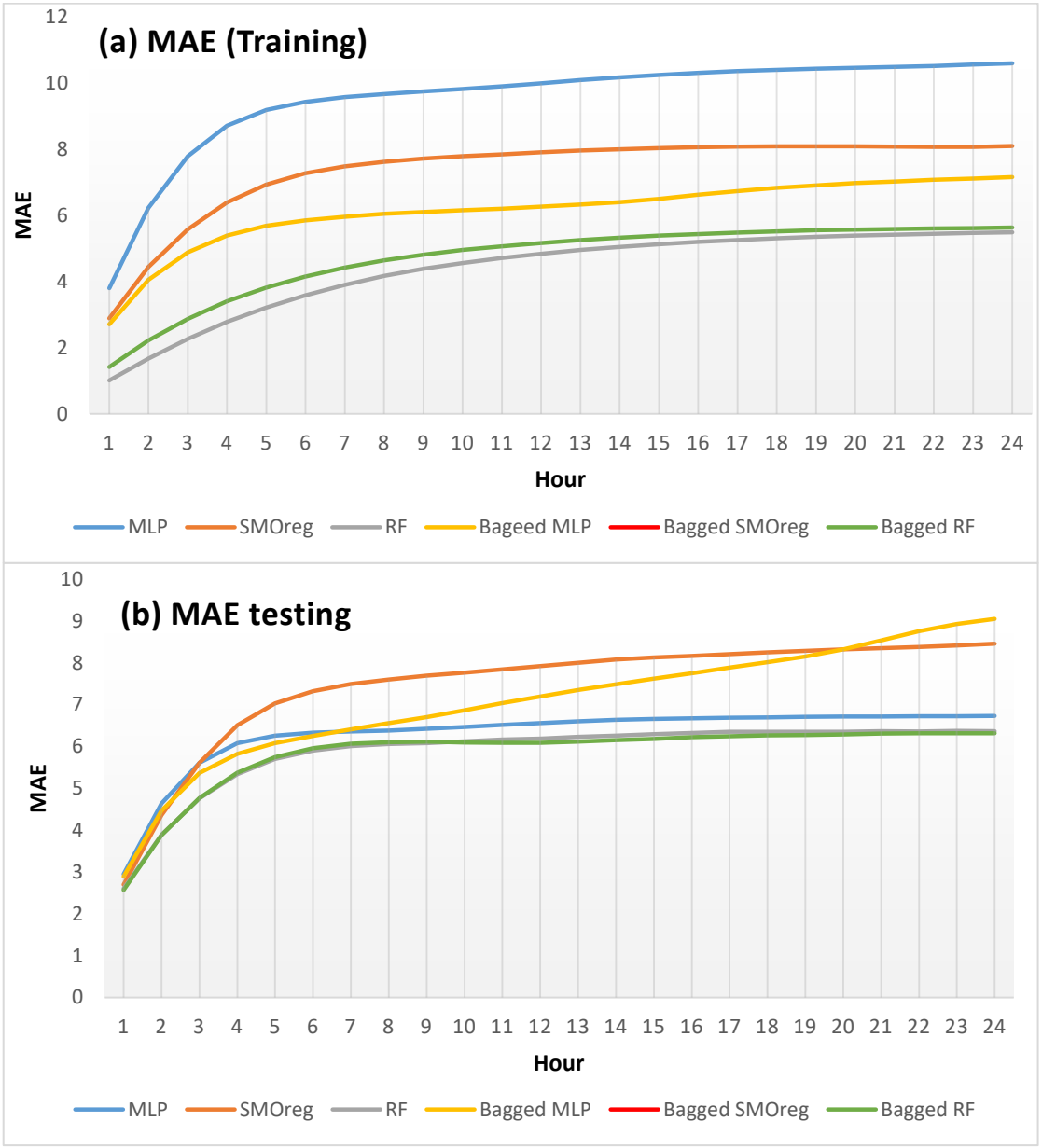

Figure 7.12: Multivariate Model (6 years' period) - MAE results for 6 classifiers (a) Training, (b) Testing

Figure 7.13 illustrates the above performances in terms of the RAE. When evaluated within the training set all forecasted values are better than a possible direct replacement from the previous 24hour period as discussed in Section 7.6.1. However when evaluated within the separate test set, only predictions beyond the third hour is better than the direct replacements from the previous day.

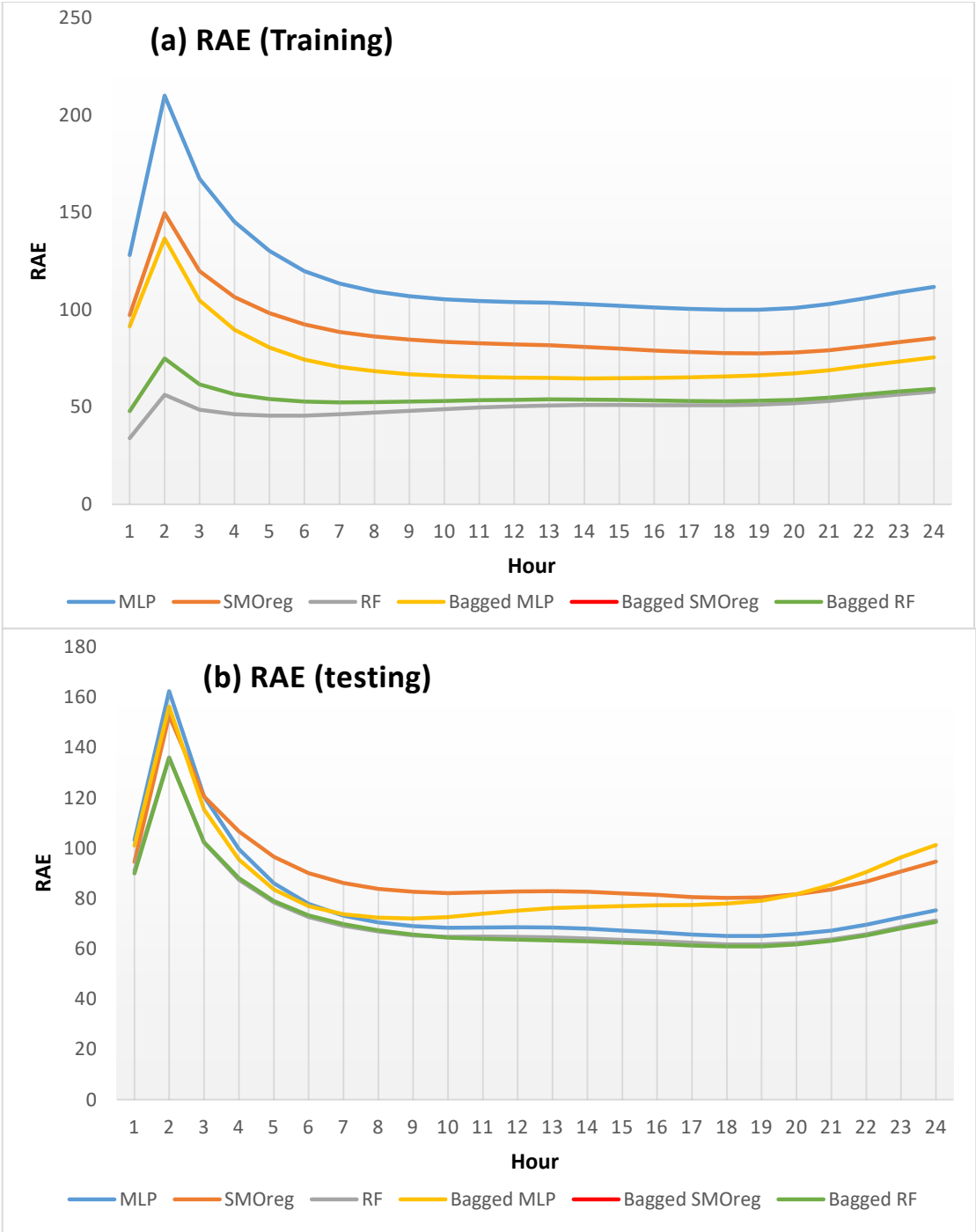

Figure 7.13: Multivariate Model (6 years' period) - RAE results for 6 classifiers (a) Training, (b) Testing

7.7 Conclusion

The research conducted within context of this chapter carried out a rigorous analysis of using machine learning algorithms including the popular single learning algorithms such as Neural Networks (Multi-Layer Perceptron), Support Vector Machines and ensemble learning algorithms such as Random Forests and Bagging

for time series prediction of ground level ozone concentrations. Both univariate and multivariate forecasting was conducted using six machine learning algorithms namely MLP, SVM (SMOreg), RF, bagged MLP, bagged SVM and bagged RF. The models were developed based on a clean, short dataset of approximately 7 months long and a long term dataset spanning approximately five years, with filled missing values. 90% of a considered dataset was used for model building and the evaluations were done on the part of the dataset used for training (i.e. model building) and also on the separated out 10% dataset. It was found that the suitability of different algorithms for forecasting varied depending on the size of the dataset, whether used for univariate or multivariate forecasting, where the evaluations were carried out (within the training set or the separated out test set).

It can be concluded that multivariate forecasting is far more accurate than univariate forecasting. This is expected as in multivariate forecasting, in addition to exploiting the univariate properties of the target variable (ozone concentration), the proposed approach also considers the impact on other parameters that directly impact the target variable such as concentrations of gases such as NO_x which is the gas that decomposes due to natural phenomena, creating ozone, and on meteorological parameters such as temperature, solar radiation and wind speed etc. Thus more accurate predictions to the future can be performed using multivariate forecasting.

The experiments conducted also revealed that when using the bagged MLP for forecasting, it is important that the correct number of bags is decided. If not the impact of Bagging on MLP can be either marginal or even detrimental. For the dataset used in the proposed experiments 15 bags were found to be an optimal number.

Bagged versions of the machine learning algorithms MLP, SMOreg and RF generally performed better than the single classifier versions. It is also observed that in general when the model created is tested on a new dataset, the forecasting errors noticeably increases. Further when the training dataset is smaller (i.e. over a shorter period of time), the model created is less accurate as long term ‘trends’ are not learnt.

Overall the best performance was indicated by RF and bagged RF. This is due to the fact that RF itself is an ensemble learning algorithm as compared to the single

classifier learning algorithms MLP and SMOreg. When RF is used as the base classifier of Bagging, the results are further improved.

Finally, it can be concluded that both univariate and multivariate forecasting of ground level ozone concentrations can be carried out successfully, in particular by ensemble learning algorithms.

CHAPTER 8

Conclusion and Future Work

The research presented in this thesis investigated the use of a comprehensive number of machine learning algorithms in the modelling, prediction and forecasting of ground level ozone. Within the literature review conducted within the research found that in the area of Environmental Pollution analysis, the modelling and prediction techniques widely used are the more popular, traditional, learning algorithms that are based on, Artificial Neural Networks, Support Vector Machines etc. Instead of using such single Learning Algorithms the use of ensemble learning algorithms (meta learning algorithms) promises more accurate modelling of complex data and hence can result in better prediction accuracies. The purpose of the research conducted in this thesis was to investigate this observation in detail. To this effect the predication and forecasting of the ground level ozone using ensemble learning approaches has been examined using two different data sets, the Sohar dataset provided by Sohar University, Sultanate of Oman and DEFRA dataset provided by the Department for Environment Food & Rural Affairs, UK. The Sohar dataset has been used to model spatial relationships between the ozone concentration and concentrations of other gases known to create ozone and a number of meteorological parameters. The models created thus were used to predict ozone concentration variations that occur due to variations of the abovementioned concentrations and parameters. On the other hand the DEFRA dataset has been used to conduct a time-series analysis, forecasting ozone concentrations based both on univariate and multivariate analysis.

The three main contributions of the thesis were presented in Chapters 5,6, and 7, where the key focus was to model ground level ozone concentration using machine learning approaches. The thesis has been organized to present each contribution in a separate chapter. The first contribution, Chapter 5, presented a comprehensive investigation into employing the state of the art machine learning techniques, single base classifiers and meta learning classifiers in the modelling and prediction of

ground level ozone concentration, based on concentrations of other common pollutants and environmental parameters. The research presented proves the superiority of three meta learning classifiers, namely, Bagging, Stacking and Voting, over widely used single classifiers such as Artificial Neural Networks, Support Vector Machines etc., in the prediction of atmospheric pollution. In the second contributory chapter, Chapter 6, the use of the ensemble learning algorithm Bagging was investigated in detail. To enhance the performance of the ensemble learning algorithm, Bagging, further, feature/attribute selection and model parameter optimisations were carried out. The results proved the ability of such additional adjustments to further improve the accuracy of prediction already provided by the ensemble learning algorithms. The third contributory chapter focused on time series analysis and forecasting of ozone concentrations, 24 hours to the future at hourly intervals. Both univariate and multivariate forecasting was conducted. The multivariate forecasting provided significantly improved levels of accuracy as compared to univariate analysis. Out of the machine learning algorithms investigated, RF and Bagging with RF provided best results under most experimental conditions. Detailed analysis and conclusions of this research were provided in Chapter 7.

Overall the research conducted in this thesis concludes that ensemble learning algorithms provide superior performance in modelling ozone concentration variations and can be used to effectively predict ozone concentrations based on parameters and concentration of other gases known to impact ozone formation and forecast ozone concentrations of the future, 24 hrs ahead, using both univariate and multivariate analysis. The performance accuracy obtained by ensemble learning algorithms are much better than the level of accuracy obtained by the popularly and widely used machine learning algorithms in atmospheric pollution modelling and prediction, i.e. Neural Networks and Support Vector machines. Therefore, it is recommended that these approaches are used in future to replace the standard approaches used widely in literature.

8.1. Future Work

This research has comprehensively investigated modelling ozone concentration using the state of art machine learning algorithms and compared them with using ensemble

learning algorithms. The research has provided an intensive analysis of modeling and forecasting. However, one outstanding issue where it is possible to extend this research further is the sensitivity analysis of the algorithms. Sensitivity analysis can be carried out using partial derivative methods to examine the sensitivity of learning algorithms to input. There have been several studies which reviews the sensitivity analysis and its different objectives. [103],[104] have presented a review of sensitive analysis in the area of environment pollution monitoring. Most of the sensitivity analysis approaches presented in literature shared the concept of varying one parameter at a time while the other parameters are maintained fixed. However, some other studies in environmental modelling have looked on the relationship between the independent variables and the dependent variables and how it affects the model performance. As future research it can be proposed to carry out sensitivity analysis of algorithms that performed best, giving the ability to compare and contrast the performance of algorithms more accurately. However this will need a significant amount of effort which was beyond the scope of research conducted within this thesis.

Further it is recommended that the use of convolutional neural networks and deep learning be investigated. These are machine learning approaches have recently being used in a wide set of application areas producing excellent results.

References

[1] U.S. Environmental Protection Agency, "Guidelines for Developing an Air Quality (ozone and PM2.5) Forecasting Program," 2003.

[2] Department of the Environment and Heritage, "Ground-level ozone (O3) - Air quality fact sheet," *Commonwealth of Australia*, 2005. [Online]. Available: http://www.environment.gov.au/protection/publications/factsheet-ground-level-ozone-o3. [Accessed: 05-Nov-2016].

[3] D. M. Agudelo–Castaneda, E. C. Teixeira, and F. N. Pereira, "Time–series analysis of surface ozone and nitrogen oxides concentrations in an urban area at Brazil," *Atmospheric Pollution Research*, vol. 5, pp. 411–420, 2014.

[4] M. Jerrett, R. T. Burnett, A. P. I. C, K. Ito, G. Thurston, D. Krewski, Y. Shi, E. Calle, and M. Thun, "Ozone exposure and mortality.," *The New England journal of medicine*, vol. 360, p. 2788; author reply 2788-2789, 2009.

[5] M. Lippmann, Ed., *Environmental Toxicants*, 3 rd. Hoboken, NJ, USA, NJ, USA: John Wiley & Sons, Inc., 2009.

[6] A. Nawahda, K. Yamashita, T. Ohara, J. Kurokawa, and K. Yamaji, "Evaluation of Premature Mortality Caused by Exposure to PM2.5 and Ozone in East Asia: 2000, 2005, 2020," *Water, Air, & Soil Pollution*, vol. 223, no. 6, pp. 3445–3459, Mar. 2012.

[7] WHO, "Health risks of particulate matter from long-range transboundary air pollution," *Pollution Atmospherique*, no. 190, p. 169, 2006.

[8] A. Nawahda, "An assessment of adding value of traffic information and other attributes as part of its classifiers in a data mining tool set for predicting surface ozone levels," *Process Safety and Environmental Protection*, vol. 99, pp. 149–158, Jan. 2016.

[9] K. P. Singh, S. Gupta, and P. Rai, "Identifying pollution sources and predicting urban air quality using ensemble learning methods," *Atmospheric Environment*, vol. 80, pp. 426–437, Dec. 2013.

[10] A. J. Cannon and E. R. Lord, "Forecasting Summertime Surface-Level Ozone Concentrations in the Lower Fraser Valley of British Columbia: An Ensemble Neural Network Approach," *Journal of the Air & Waste Management Association*, vol. 50, no. 3, pp. 322–339, Mar. 2000.

[11] I. H. Witten, E. Frank, and M. Hall, *Data Mining Practical Machine Learning Tools and Techniques*, 3rd ed. Elsevier, 2011.

[12] WEKA; the University of Waikato, "Weka 3 - Data Mining with Open Source Machine Learning Software in Java." [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/index.html. [Accessed: 27-Feb-2015].

[13] T. G. Dietterich, "Ensemble Methods in Machine Learning," *Lecture Notes in Computer Science*, vol. 1857, pp. 1–15, 2000.

[14] F. C. Morabito and M. Versaci, "Fuzzy neural identification and forecasting

techniques to process experimental urban air pollution data.," *Neural networks : the official journal of the International Neural Network Society*, vol. 16, no. 3–4, pp. 493–506, Jan. 2003.

[15] S. McKeen, J. Wilczak, G. Grell, I. Djalalova, S. Peckham, E.-Y. Hsie, W. Gong, V. Bouchet, S. Menard, R. Moffet, J. McHenry, J. McQueen, Y. Tang, G. R. Carmichael, M. Pagowski, A. Chan, T. Dye, G. Frost, P. Lee, and R. Mathur, "Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004," *Journal of Geophysical Research*, vol. 110, no. D21, p. D21307, 2005.

[16] L. D. Monache and R. B. Stull, "An ensemble air-quality forecast over western Europe during an ozone episode," *Atmospheric Environment*, vol. 37, no. 25, pp. 3469–3474, 2003.

[17] M. van Loon, R. Vautard, M. Schaap, R. Bergström, B. Bessagnet, J. Brandt, P. J. H. Builtjes, J. H. Christensen, C. Cuvelier, A. Graff, J. E. Jonson, M. Krol, J. Langner, P. Roberts, L. Rouil, R. Stern, L. Tarrasón, P. Thunis, E. Vignati, L. White, and P. Wind, "Evaluation of long-term ozone simulations from seven regional air quality models and their ensemble," *Atmospheric Environment*, vol. 41, no. 10, pp. 2083–2097, 2007.

[18] V. Mallet and B. Sportisse, "Ensemble-based air quality forecasts: A multimodel approach applied to ozone," *Journal of Geophysical Research*, vol. 111, no. D18, p. D18302, 2006.

[19] W. Wang, C. Men, and W. Lu, "Online prediction model based on support vector machine," *Neurocomputing*, vol. 71, no. 4–6, pp. 550–558, Jan. 2008.

[20] S. . Abdul-Wahab and S. . Al-Alawi, "Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks," *Environmental Modelling & Software*, vol. 17, no. 3, pp. 219–228, 2002.

[21] U. Schlink, S. Dorling, E. Pelikan, G. Nunnari, G. Cawley, H. Junninen, A. Greig, R. Foxall, K. Eben, T. Chatterton, J. Vondracek, M. Richter, M. Dostal, L. Bertucco, M. Kolehmainen, and M. Doyle, "A rigorous inter-comparison of ground-level ozone predictions," *Atmospheric Environment*, vol. 37, no. 23, pp. 3237–3253, Jul. 2003.

[22] M. S. Baawain and A. S. Al-Serihi, "Systematic approach for the prediction of ground-level air pollution (around an industrial port) using an artificial neural network," *Aerosol and Air Quality Research*, vol. 14, pp. 124–134, 2014.

[23] N. Loya, I. Olmos Pineda, D. Pinto, H. Gómez-Adorno, and Y. Alemán, "Forecast of air quality based on ozone by decision trees and neural networks," in *Mexican International Conference on Artificial Intelligence*, 2012, pp. 97–106.

[24] B. Mileva-Boshkoska and M. Stankovski, "Prediction of missing data for ozone concentrations using support vector machines and radial basis neural networks," *Informatica*, vol. 31, no. 4, 2007.

[25] J. Gomez-Sanchis, J. Martın-Guerrero, E. Soria-Olivas, J. Vila-France´s, J. Carrasco, and S. Valle-Tasco´n, "Neural networks for analysing the relevance of input variables in the prediction of tropospheric ozone concentration," *Atmospheric Environment*, vol. 40, no. 32, pp. 6173–6180, Oct. 2006.

[26] A. S. Luna, M. L. L. Paredes, G. C. G. de Oliveira, and S. M. Corrêa, “Prediction of ozone concentration in tropospheric levels using artificial neural networks and support vector machine at Rio de Janeiro, Brazil,” *Atmospheric Environment*, vol. 98, pp. 98–104, 2014.

[27] A. Coman, A. Ionescu, and Y. Candau, “Hourly ozone prediction for a 24-h horizon using neural networks,” *Environmental Modelling & Software*, vol. 23, no. 12, pp. 1407–1421, Dec. 2008.

[28] W.-Z. Lu and D. Wang, “Learning machines: Rationale and application in ground-level ozone prediction,” *Applied Soft Computing`*, vol. 24, pp. 135–141, Nov. 2014.

[29] W. Z. Lu, H. Y. Fan, and S. M. Lo, “Application of evolutionary neural network method in predicting pollutant levels in downtown area of Hong Kong,” *Neurocomputing*, vol. 51, pp. 387–400, Apr. 2003.

[30] D. Wang and W.-Z. Lu, “Interval estimation of urban ozone level and selection of influential factors by employing automatic relevance determination model.,” *Chemosphere*, vol. 62, no. 10, pp. 1600–11, Mar. 2006.

[31] S. Sousa, F. Martins, M. Alvim-Ferraz, and M. Pereira, “Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations,” *Environmental Modelling & Software*, vol. 22, no. 1, pp. 97–103, Jan. 2007.

[32] A. Sotomayor-Olmedo, M. A. Aceves-Fernández, E. Gorrostieta-Hurtado, C. Pedraza-Ortega, J. M. Ramos-Arreguín, and J. E. Vargas-Soto, “Forecast Urban Air Pollution in Mexico City by Using Support Vector Machines: A Kernel Performance Approach,” *International Journal of Intelligence Science*, vol. 3, no. 3, pp. 126–135, Jul. 2013.

[33] M. Riga, F. A. Tzima, K. Karatzas, and P. A. Mitkas, “Development and Evaluation of Data Mining Models for Air Quality Prediction in Athens, Greece,” in *Information Technologies in Environmental Engineering*, 2009, pp. 331–344.

[34] E. B. Ballester, G. C. i Valls, J. . Carrasco-Rodriguez, E. S. Olivas, and S. del Valle-Tascon, “Effective 1-day ahead prediction of hourly surface ozone concentrations in eastern Spain using linear models and neural networks,” *Ecological Modelling*, vol. 156, no. 1, pp. 27–41, 2002.

[35] K. B. Shaban, A. Kadri, and E. Rezk, “Urban Air Pollution Monitoring System With Forecasting Models,” *IEEE Sensors Journal*, vol. 16, no. 8, pp. 2598–2606, Apr. 2016.

[36] M. Gardner and S. R. Dorling, “Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London,” *Atmospheric Environment*, vol. 33, no. 5, pp. 709–719, Feb. 1999.

[37] A. Russo and A. O. Soares, “Hybrid Model for Urban Air Pollution Forecasting: A Stochastic Spatio-Temporal Approach,” *Mathematical Geosciences*, vol. 46, no. 1, pp. 75–93, Jan. 2014.

[38] C. Capilla, “Multilayer perceptron and regression modelling to forecast hourly nitrogen dioxide concentrations,” *WIT Transactions on Ecology and the Environment*, vol. 183, pp. 39–48, 2014.

[39] S. F. Crone and R. Dhawan, “Forecasting Seasonal Time Series with Neural Networks: A Sensitivity Analysis of Architecture Parameters,” in *International Joint Conference on Neural Networks*, 2007, pp. 2099–2104.

[40] K. P. Singh, S. Gupta, A. Kumar, and S. P. Shukla, “Linear and nonlinear modeling approaches for urban air quality prediction.,” *The Science of the total environment*, vol. 426, pp. 244–55, Jun. 2012.

[41] P.-W. G. Liu, J.-H. Tsai, H.-C. Lai, D.-M. Tsai, and L.-W. Li, “Establishing multiple regression models for ozone sensitivity analysis to temperature variation in Taiwan,” *Atmospheric Environment*, vol. 79, pp. 225–235, Nov. 2013.

[42] N. Jiang and M. L. Riley, “Exploring the Utility of the Random Forest Method for Forecasting Ozone Pollution in SYDNEY,” *Journal of Environment Protection and Sustainable Development*, vol. 1, no. 5, pp. 245–254, 2015.

[43] P. Yang, Y. Hwa Yang, B. B. Zhou, and A. Y. Zomaya, “A Review of Ensemble Methods in Bioinformatics,” *Current Bioinformatics*, vol. 5, no. 4, pp. 296–308, 2010.

[44] A. C. Tan and D. Gilbert, “Ensemble machine learning on gene expression data for cancer classification.” University of Glasgow, 2003.

[45] G. Liang and C. Zhang, “Empirical Study of Bagging Predictors on Medical Data,” in *Proceedings of the Ninth Australasian Data Mining Conference*, 2011, vol. 121, pp. 31–40.

[46] E. Alfaro, N. García, M. Gámez, and D. Elizondo, “Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks,” *Decision Support Systems*, vol. 45, no. 1, pp. 110–122, Apr. 2008.

[47] L. A. Gabralla and A. Abraham, “Prediction of Oil Prices Using Bagging and Random Subspace,” in *Proceedings of the Fifth International Conference on Innovations in Bio-Inspired Computing and Applications IBICA 2014*, 2014, pp. 343–354.

[48] A. Fathima, J. A. Mangai, and B. B. Gulyani, “An ensemble method for predicting biochemical oxygen demand in river water using data mining techniques,” *International Journal of River Basin Management*, vol. 12, no. 4, pp. 357–366, Oct. 2014.

[49] P. Tüfekci, “Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods,” *International Journal of Electrical Power & Energy Systems*, vol. 60, pp. 126–140, Sep. 2014.

[50] Y. Feng, W. Zhang, D. Sun, and L. Zhang, “Ozone concentration forecast method based on genetic algorithm optimized back propagation neural networks and support vector machine data classification,” *Atmospheric Environment*, vol. 45, no. 11, pp. 1979–1985, Apr. 2011.

[51] J. Mendes-Moreira, C. Soares, A. M. Jorge, and J. F. De Sousa, “Ensemble approaches for regression: A survey,” *ACM Computing Surveys*, vol. 45, no. 1, pp. 1–40, Nov. 2012.

[52] S. B. Kotsianti and D. Kanellopoulos, “Combining Bagging , Boosting and Dagging

for Classification Problems," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems. Springer Berlin Heidelberg*, 2007, pp. 493–500.

[53] N. Rooney, D. Patterson, S. Anand, and A. Tsymbal, "Dynamic Integration of Regression Models," in *International Workshop on Multiple Classifier Systems*, 2004, pp. 164–173.

[54] OPSIS, "UV DOAS Technique," 2014. [Online]. Available: http://opsis.se/Techniques/UVDOASTechnique/tabid/632/Default.aspx. [Accessed: 09-Feb-2015].

[55] WikiLectures, "Lambert-Beer's law," 2014. [Online]. Available: http://www.wikilectures.eu/index.php/Lambert-Beer's_law. [Accessed: 03-Feb-2015].

[56] H. Akaike, "Information Theory and an Extension of the Maximum Likelihood Principle," in *Selected Papers of Hirotugu Akaike*, Springer New York, 1998, pp. 199–213.

[57] S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, and K. K. Murthy, "Improvements to the SMO algorithm for SVM regression.," *IEEE transactions on neural networks*, vol. 11, no. 5, pp. 1188–1193, Jan. 2000.

[58] Y. Wang and I. H. Witten, "Induction of model trees for predicting continuous classes," in *Proceedings of the Poster Papers of the European Conference on Machine Learning*, 1996.

[59] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-Based Learning Algorithms," *Machine Learning*, vol. 6, no. 1, pp. 37–66, Jan. 1991.

[60] W. Iba and P. Langley, "Induction of One-Level Decision Trees," in *Proceedings of the ninth international conference on machine learning*, 1992, pp. 233–240.

[61] R. Kohavi, "The Power of Decision Tables," in *European conference on machine learning*, 1995, pp. 174–189.

[62] G. Holmes, M. Hall, and E. Frank, "Generating Rule Sets from Model Trees," in *Australasian Joint Conference on Artificial Intelligence*, 1999, pp. 1–12.

[63] J. Fürnkranz, "Separate-and-Conquer Rule Learning," *Artificial Intelligence Review*, vol. 13, no. 1, pp. 3–54, 1999.

[64] E. Frank, M. Hall, and B. Pfahringer, "Locally weighted naive bayes," in *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, 2002, pp. 249–256.

[65] L. E. T. John G. Cleary, "K*: An Instance-based Learner Using an Entropic Distance Measure," in *Proceedings of the 12th International Conference on Machine learning*, 1995, vol. 5, pp. 108–114.

[66] H. I. Erdal and O. Karakurt, "Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms," *Journal of Hydrology*, vol. 477, pp. 119–128, Jan. 2013.

[67] D. Hern, G. Mart, D. Hernandez-Lobato, G. Martinez-Munoz, and A. Suarez, "Pruning in Ordered Regression Bagging Ensembles," in *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, 2006, pp. 1266–1273.

[68] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[69] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[70] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[71] Tin Kam Ho, "The random subspace method for constructing decision forests," *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[72] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, Feb. 2002.

[73] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, Jan. 1992.

[74] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *International Joint Conference on Artificial Intelligence*, vol. 14, no. 12, pp. 1137–1143, 1995.

[75] Wikipedia, "Seymour Geisser." [Online]. Available: https://en.wikipedia.org/wiki/Seymour_Geisser. [Accessed: 24-Jul-2016].

[76] MathsIsFun, "Correlation," 2016. [Online]. Available: http://www.mathsisfun.com/data/correlation.html. [Accessed: 04-Apr-2016].

[77] R. Nau, "Forecasting with moving averages," 2014. [Online]. Available: http://people.duke.edu/~rnau/Notes_on_forecasting_with_moving_averages--Robert_Nau.pdf. [Accessed: 08-Sep-2016].

[78] C. J. Willmott, "Some Comments on the Evaluation of Model Performance," *Bulletin of the American Meteorological Society*, vol. 63, no. 11, pp. 1309–1313, Nov. 1982.

[79] J. Brownlee, "Feature Selection to Improve Accuracy and Decrease Training Time," *Machine Learning Mastery*, 2014. [Online]. Available: http://machinelearningmastery.com/feature-selection-to-improve-accuracy-and-decrease-training-time/. [Accessed: 27-Jul-2016].

[80] M. A. Hall and G. Holmes, "Benchmarking Attribute Selection Techniques for Discrete Class Data Mining," *IEEE transactions on knowledge and data engineering*, vol. 15, no. 6, pp. 1437–1447, 2003.

[81] Univeristy of Waikato, "Optimizing parameters," *the Univeristy of Waikato*, 2016. [Online]. Available: https://weka.wikispaces.com/Optimizing+parameters.

[82] A. Plaia and A. L. Bondì, "Single imputation method of missing values in environmental pollution data sets," *Atmospheric Environment*, vol. 40, no. 38, pp.

7316–7330, 2006.

[83] Y. Kim and E. Riloff, “Stacked Generalization for Medical Concept Extraction from Clinical Notes,” in *Workshop on Biomedical Natural Language Processing*, 2015.

[84] S. Diplaris, G. Tsoumakas, P. a. Mitkas, and I. Vlahavas, “Protein classification with multiple algorithms,” in *Panhellenic Conference on Informatics*, 2005, pp. 448–456.

[85] M. Eric, U. C. Berkeley, J. Scott, U. C. Berkeley, E. M. Burger, and S. J. Moura, “Building Electricity Load Forecasting via Stacking Ensemble Learning Method with Moving Horizon Optimization,” 2015.

[86] P. K. Mahato and V. Attar, “Prediction of gold and silver stock price using ensemble models,” in *2014 International Conference on Advances in Engineering & Technology Research (ICAETR - 2014)*, 2014, pp. 1–4.

[87] P. Panov and S. Džeroski, “Combining Bagging and Random Subspaces to Create Better Ensembles,” in *International Symposium on Intelligent Data Analysis*, 2007, pp. 118–129.

[88] T. Windeatt, “Ensemble MLP classifier design,” *Studies in Computational Intelligence*, vol. 137, pp. 133–147, 2008.

[89] “Stability (learning theory) - Wikipedia, the free encyclopedia.” [Online]. Available: https://en.wikipedia.org/wiki/Stability_(learning_theory). [Accessed: 25-Oct-2015].

[90] O. Bousquet and A. Elisseeff, “Stability and Generalization,” *Journal of Machine Learning Research*, vol. 2, pp. 499–526, 2002.

[91] G. Valentini, M. Muselli, and F. Ruffino, “Cancer recognition with bagged ensembles of support vector machines,” *Neurocomputing*, vol. 56, pp. 461–466, 2004.

[92] H. Hsu, P. A. Lachenbruch, H. Hsu, and P. A. Lachenbruch, “Paired T Test,” in *Wiley Encyclopedia of Clinical Trials*, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2008.

[93] S.-Q. Wang, J. Yang, and K.-C. Chou, “Using stacked generalization to predict membrane protein types based on pseudo-amino acid composition.,” *Journal of theoretical biology*, vol. 242, no. 4, pp. 941–6, Oct. 2006.

[94] M. I. R. Caffé, P. S. Perez, and J. A. Baranauskas, “Evaluation of Stacking on Biomedical Data *,” *Journal of Health Informatics*, vol. 4, no. 3, pp. 67–72, 2012.

[95] Pentaho-A hitachi Group Company, “Time Series Analysis and Forecasting with Weka.” [Online]. Available: http://wiki.pentaho.com/display/DATAMINING/Time+Series+Analysis+and+Forecasting+with+Weka. [Accessed: 09-Dec-2015].

[96] D. Gerbing, “Time Series Components,” *School of Business Administration, Portland State University*, 2016.

[97] C. Napagoda, “Web Site Visit Forecasting Using Data Mining Techniques,” *International Journal of Scientific & Technology Research*, vol. 2, no. 12, pp. 170–174, 2013.

[98] M. Reinstadler, M. Braunhofer, M. Elahi, and F. Ricci, “Predicting parking lots

occupancy in Bolzano," *Academic Project, Computer Science, Free University of Bolzano Italy, Bolzano*, 2013.

[99] N. K. Ahmed, A. F. Atiya, N. El Gayar, and H. El-Shishiny, "An empirical comparison of machine learning models for time series forecasting," *Econometric Reviews*, vol. 29, no. 5–6, pp. 594–621, 2010.

[100] J. H. Cochrane, "Time Series for Macroeconomics and Finance," *Graduate School of Business, University of Chicago*, 1997.

[101] R. K. R. A. Agrawal, "An Introductory Study on Time Series Modeling and Forecasting," *arXiv preprint arXiv:1302.6613.*, pp. 1–68, 2013.

[102] P. Cortez, "Sensitivity analysis for time lag selection to forecast seasonal time series using Neural Networks and Support Vector Machines," in *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN 2010)*, 2010, pp. 3694–3701.

[103] D. M. Hamby, "A review of techniques for parameter sensitivity analysis of environmental models," *Environmental Monitoring and Assessment*, vol. 32, no. 2, pp. 135–154, Sep. 1994.

[104] F. Pianosi, K. Beven, J. Freer, J. W. Hall, J. Rougier, D. B. Stephenson, and T. Wagener, "Sensitivity analysis of environmental models: A systematic review with practical workflow," *Environmental Modelling & Software*, vol. 79, pp. 214–232, 2016.

Appendix-A

Scholarly Contribution

This works has resulted three papers which has been submitted/published as conference or journal papers. The details could be found in below sections:

I. Conference Published

Al Abri, E.S., Edirisinghe, E.A. and Nawadha, A., 2015. Modelling ground-level ozone concentration using ensemble learning algorithms. Proceedings of the International Conference on Data Mining (DMIN), 27th-30th July 2015, Las Vegas, USA, pp.148-154.

II. Journal Paper under Review

Al Abri, E.S., Edirisinghe, E.A., Nawadha,A., and Dawson, C.W., The Use of Meta-Learning Ensemble Algorithms for the Prediction of Ground-Level Ozone, submitted to Big Data Research, Elsevier , May 2016

III. Conference Paper will be Submitted

Al Abri, E.S., Edirisinghe, E.A., and Dawson, C.W., Ability of ensemble learning to forecast ground level ozone concentration, will be submitted to 13th International Conference on Machine Learning and Data Mining MLDM 2017