Vehicular Emissions Prediction with CART-BMARS Hybrid Models

S. D. Oduro[a], Q.P. Ha[a], H. Duc[b]

[a]*Faculty of Engineering & IT, University of Technology Sydney, Australia*
[b]*Office of Environment & Heritage, Sydney, Australia*

Abstract

Vehicular emission models play a key role in the development of reliable air quality modelling systems. To minimise uncertainties associated with these models, it is essential to match the high-resolution requirements of emission models with up-to-date information. However, these models are usually based on average trip speed, not on environmental parameters like ambient temperature, and vehicle's motion characteristics, such as speed, acceleration, load and power. This contributes to the degradation of its predictive performance. In this paper, we propose to use the non-parametric Classification and Regression Trees (CART), the Boosting Multivariate Adaptive Regression Splines (BMARS) algorithm and a combination of them in hybrid models to improve the accuracy of vehicular emission prediction using on-board measurements and the chassis dynamometer testing. The experimental comparison between the proposed CART-BMARS hybrid model with the BMARS and artificial neural networks (ANNs) algorithms demonstrates its effectiveness and efficiency in estimating vehicular emissions.

Keywords: Vehicular emissions, on-board emission measurement, chassis dynamometer testing, CART-BMARS, ANNs.

1. Introduction

Poor air quality has become a serious problem in recent years in many cities and their surrounding areas due to increasing population, motor vehicles and industries. To be environmentally-sustainable, efforts have been made to improve energy efficiency and to reduce air pollutant emissions in both generation and consumption sides [1]. Among air pollutants coming

from all sources, anthropogenic emissions have been the main concern in air-quality modelling and control. The problem is exacerbated as the world demand of transport is projected to increase by 45% by the year 2030 [2] while the steady growth in vehicular population in the urban areas. This will involve the increase in the number of motor vehicles and consequently the emissions impact. As vehicular emissions are produced at the ground level, they have harmful effects directly on the reception population [3].

It is a fact that the transport sector is growing quickly and providing convenient and quick access to any geographical location. However, it also brings disadvantages like noise, congestion and pollutant emissions such as carbon monoxide (CO), nitrogen oxides (NO_X), total volatile hydrocarbon (THC), Carbon dioxide (CO_2), which are primarily responsible for global warming ([4], [5]). The amount of CO_2 emitted from distance traveled is directly proportional to fuel economy with every litre of gasoline burned releasing about 2.4 kg of CO_2 [6]. The problem of vehicular emissions becomes more severe when the traffic flow is congested or interrupted especially when the delays and disruptions occur frequently. These phenomena are regularly observed at traffic intersections, junctions, and at signalized roadways, where traffic related characteristics combined with road and vehicle conditions contribute to the level of emissions.

Many research initiatives have been undertaken to model and predict the complexity of vehicle emissions in order to control transport air pollution [7]. However, the mechanisms by which they affect the atmosphere and degrade the urban air quality are not completely identified. Consequently, the need of comprehensive and accurate models for vehicle emissions is essential to safeguard the urban air quality, to recognize any potential changes in the climate, and to justify imposing new regulations. It is vital to increase the ability of policy-makers to reach sound and reasonable decisions about vehicle emissions and air quality in order to maintain environmental sustainability.

Air quality models are indispensable tools to assess the impact of air pollutants on human health and the urban development. The most critical part of assessment studies is to know the present as well as future air quality levels. In this paper, we aim to improve the prediction accuracy of emissions modelling based on data collected from chassis dynamometer and on-board measurement systems. The dynamometer testing is one of the three typical vehicle tailpipe emission measurement methods, where emissions from vehicles are measured under laboratory conditions during a driving cycle to simulate vehicle road operations [8]. The real world on-board emissions

measurement is widely recognised as a desirable approach for quantifying emissions from vehicles since data are collected under real-world conditions at any location travelled by the vehicle [9]. Using on-board measurements, variability in traffic emissions as a result of changes in roadway characteristics, vehicle's location and operation mode, driver, or other factors can be represented and analysed more reliably than with the other methods [10]. This is because measurements are obtained during real world driving, eliminating the concern about non-representativeness that is often an issue with dynamometer testing, and at any location, eliminating the setting restrictions inherent in remote sensing. Though the on-board measuring technique seems to be more promising, the need to improve the prediction accuracy of emission factor by using effective statistical techniques is important in any emissions modelling approach.

Therefore, to adequately model traffic emissions, the Multivariate Adaptive Regression Splines (MARS) technique has proven to be promising [11]. However, the infulential factors such as vehicles' speed, acceleration, load, power and ambient temperature have not been fully considered therein. To enhance the prediction performance taking into account these emissions factors, in this paper we focus on integrating the Classification and Regression Trees (CART) technique with Boosting Multivariate Adaptive Regression Splines (BMARS) to provide a regression tree to better predict these continuous dependent variables for the regression model with BMARS [12]. Here, our purpose is to achieve highly-accurate estimates from the emission models from the dynamometer testing and the on-board measuring data. The effectiveness of the proposed approach is then determined by grouping the data into two parts, one for building the model (learning) and the other for validating the model (testing).

Among machine learning methods, artificial neural networks (ANNs), in particular, the multilayer feedforward networks with the back-propagation algorithm, have been widely applied in the last decades to environmental modelling [13], wherein good performance has been obtained for various vehicular emissions models [14–16] or prediction of air pollution profiles in a region [17, 18]. Therefore, it is worth comparing the results from the CART-BMARS hybrid model developed in this paper with those obtained by using the BMARS and MARS and ANNs techniques.

The organization of this paper is as follows. After the introduction, Section 2 presents the development of the CART, MARS, BMARS and ANNs algorithms. Section 3 includes vehicle information and data collection proce-

dure. Section 4 shows results obtained by using all the mentioned methods and dissusses on their advantages. Finally, Section 5 draws a conclusion for the paper.

2. Vehicular Emissions Models

The proposed model for vehicular emissions modelling is based on a combination of CART and MARS techniques coupled with a boosting algorithm to improve the learning performance.

2.1. CART modelling

The Classification and Regression Trees (CART) technique is a non-parametric solution approach to form classifications or regression trees depending on whether the dependent variable is categorical or numerical [19]. CART begins with the root node at the top of the tree, which contains the entire data for the training run [20]. A node in the CART model is either a terminal node, i.e. a node without children, or non-terminal node, i.e. a node with children [21]. The algorithm is intended for the building of a binary solutions tree consisting of the main splitters in CART. Here, to take into account not only the speed but also acceleration, load, power, and ambient temperature in our vehicular emissions model, a regression tree can result from the CART analysis, as shown typically in Fig. 1. Those cells that meet the condition within the nodes go to the left side while the remaining cells go to the right side.

The initial set of observations is divided into groups at the terminal nodes, or leaves, of the tree. The goal is to find a tree which allows for a good distribution of data with the lowest possible relative error of prediction. Each branch of the tree ends with one or two terminal nodes and each observation falls exactly into one terminal node, defined by a unique set of rules [22]. CART initially build an overgrown model to make sure that stopping rules do not prevent the model from extracting the correct patterns in data during the training run to prevent under-fitting. Consequently, the model is pruned back by penalizing model complexity and removing those splits that do not improve the accuracy significantly to prevent over-fitting. The tree structure represents a series of splits for different predictors, where predictor variables in the emission data are organized hierarchically, i.e. levels in the tree are representative of the variables' levels of significance. In CART, splits occur from the use of search algorithms to classify data into binary or multiple

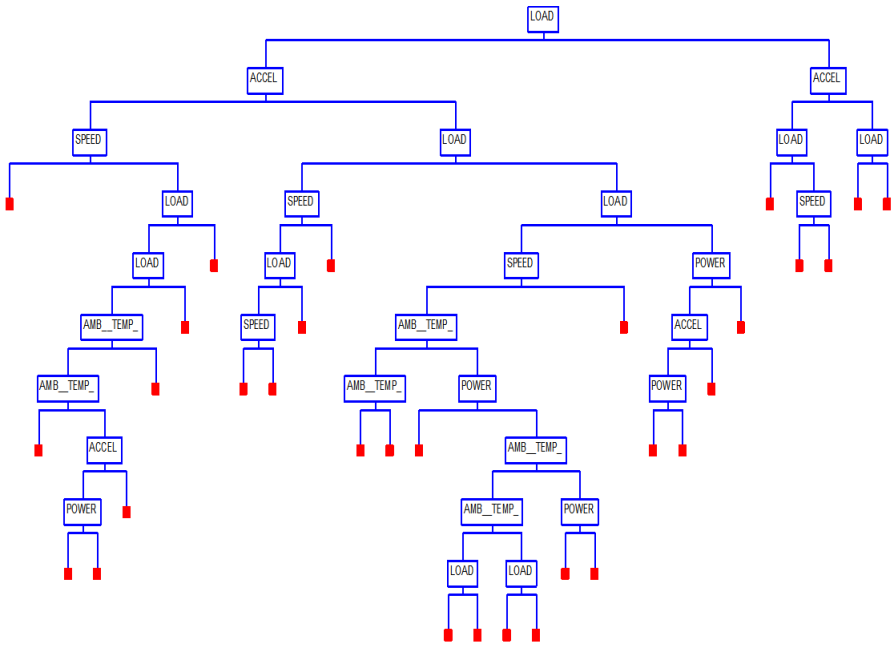


Figure 1: Regression tree from CART analysis.

classes [19] by checking all unique values across the range of data values for different predictors [23].

The CART algorithm calculates the probability (P_k) of the emission variables in the root node of the tree using relative frequencies in the entire learning data, $P_k = N_k|N; k = 1, 2, ..., K$, where N_k is the number of cells corresponding to emission variable k from the entire data N [24]. Let $P(k, t)$ denote the probability of emission variable k and $N_k(t)$ be the number of cells in node t belonging to class k, then

$$P(k, t) = P_k \times \frac{N_k(t)}{N_k}. \tag{1}$$

Now let $P(k|t)$ denote the conditional probability that the CART algortihm classifies correctly the emission variables and $P(t) = \sum_k P(k, t))$, then

$$P(k|t) = \frac{P(k, t)}{P(t)}. \tag{2}$$

In this paper, to measure the inequality among values of emission variables, we use the Gini index as a node impurity function. The splitting rule

for each unique value in the predictors is applied to find the best split of fragment data [19] from a uniform cost, i.e. the misclassification cost is equal for all classes:

$$d(t) = \sum_{k=1}^{K} \sum_{j=1}^{K-1} P(j|t)P(k|t) = \frac{1}{2}\left(1 - \sum_{k=1}^{K} P^2(k|t)\right), \tag{3}$$

or a non-uniform cost:

$$d(t) = \sum_{k=1}^{K} \sum_{j=1}^{K-1} P(j|t)P(k|t) + C(k|t), \tag{4}$$

where $C(j|k)$ represents the cost of misclassifying a cell that belongs to emission variable k into emission variable j.

To get the best split in node t, we look for the one that maximizes the node impurity function, or the misclassification cost $d(t)$, in the children of node t [24]. To make a more homogeneous subset than the previous node, the following gain function makes use of a distribution of data before and after splitting:

$$\triangle d(s,t) = d(t) - P_L d(t_L) - P_R d(t_R), \tag{5}$$

where P_L and P_R are the proportions of cells going to left node t_L (left) and right node t_R, respectively. The gain function (5) can be used to determine the goodness of a split, e.g. split s for node t [25]. A splitting value is adopted at node t to minimize the diversity obtained by the split. All the predictor data set records are assigned to one of the terminal nodes, which represent the particular class or subset of emissions variables. The training data together with this node information are supplied for MARS modelling.

2.2. MARS Modelling

The Multivariate Adaptive Regression Splines (MARS) technique, also for non-parametric regression, uses a series of basis functions to model complex (such as non-linear) relationships [26]. Its main purpose is to predict the values of a continuous dependent variable, $y(n \times 1)$, from a set of p independent explanatory variables, $X(n \times p)$, which in our case are emissions factors as mentioned above. The MARS model can be represented as:

$$y = f(X) + e, \tag{6}$$

where f is a weighted sum of basis functions that depend on X and e is an error vector of dimension $(n \times 1)$. MARS provides a greater flexibility to explore the non-linear relationship between a response variable and predictor variables by fitting the data into piecewise linear regression functions. It does not require *a priori* assumptions about the underlying functional relationship between dependent and independent variables. Instead, this relation is uncovered from a set of coefficients and piecewise polynomials of degree q basis functions (BFs) that are entirely driven from the regression data (y, X). The MARS regression model is constructed by fitting basis functions into distinct intervals of the independent variables. Generally, piecewise polynomials, also called splines, have pieces smoothly connected together. Here, the joining points of the polynomials are called knots, nodes or breakdown points, denoted by t. For a spline of degree q each segment is a polynomial function. MARS uses two-sided truncated power functions as spline basis functions, described by the following equations [27]:

$$[-(x-t)_+^q] = \begin{cases} (t-x)^q; & \text{if } x < t, \\ 0; & \text{otherwise.} \end{cases} \tag{7}$$

$$[+(x-t)_+^q] = \begin{cases} (x-t)^q; & \text{if } x > t, \\ 0; & \text{otherwise,} \end{cases} \tag{8}$$

where $q(\geq 0)$ is the power to which the splines are raised and which determines the degree of smoothness of the resultant function estimate. As an example, a pair of splines for $q = 1$ at the knot $t = 0.5$ is presented in Fig. 2.

The two-sided truncated functions of the dependent variable are basis functions that describe the underlying phenomena. The global MARS model is defined as [28]:

$$\hat{y} = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X), \tag{9}$$

where $\hat{y}$ is the predicted response; β_0 is the coefficient of the constant basis function; $h_m(X)$ is the mth basis function, which can be a single spline function or an interaction of two (or more) spline functions; β_m is the coefficient of the mth basis function; and M is the number of basis functions included in the MARS model. To fit a MARS model, three main steps are applied. In the first step, i.e., the constructive phase, basis functions are added to

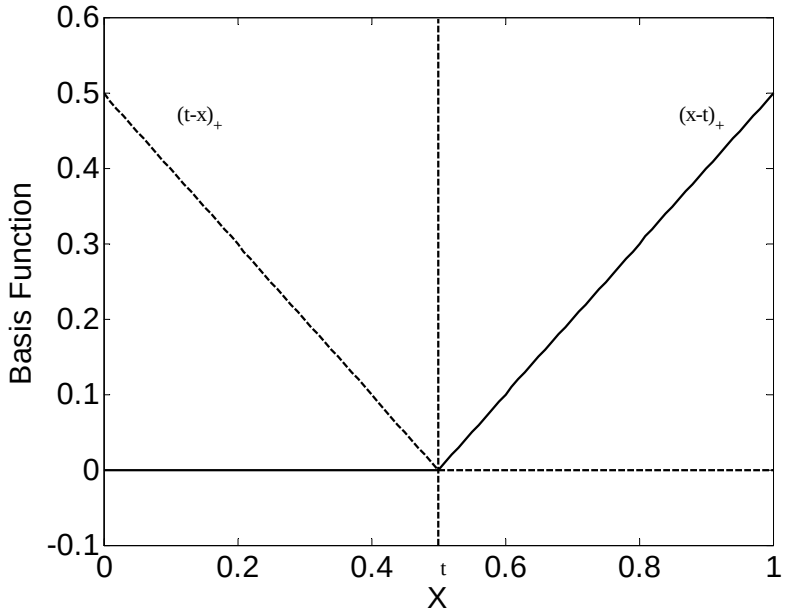


Figure 2: A graphical representation of a spline basis function.

the model using a forward stepwise procedure. The predictor and the knot location that contribute significantly to the model accuracy are selected. In this stage, interactions are also introduced to examine if they could improve the model fit. To improve the prediction, the redundant basis functions are removed one at a time using the backward stepwise procedure, in the second stage. MARS utilizes the generalized cross-validation (GVC), incorporating the criterion for finding the overall best model from a sequence of fitted models, and is estimated by the lack-of-fit [29]:

$$\text{GCV} = \frac{1}{N} \frac{\sum_{i=1}^{N} \left(y_i - \hat{f}(X_i) \right)^2}{\left[1 - \frac{\tilde{C}(M)}{N} \right]^2}, \tag{10}$$

where $\left[1 - \frac{\tilde{C}(M)}{N} \right]^2$ is a complexity function, and $\tilde{C}(M)$ is defined as $\tilde{C}(M) = C(M) + dM$, in which $C(M)$ is the number of parameters to be fit and smoothing parameter d is a user-defined cost for each basis function opti-

mization. The higher the cost d is, the more basis functions will be eliminated [28]. Finally, the third step to select the optimal MARS model, based on an evaluation of the prediction characteristics of different fitted MARS models.

2.3. BMARS modelling

Boosting has been widely-used for predictive modelling as it offers an efficient, simple technique to manipulate additive modelling [19], that can convert weak learners to potentially a strong learner, i.e. a classifier well-correlated with the true classification. A succession of models can be built iteratively from boosting. At this point, the examples are being trained and re-weighted. Finally, each model or a weak classifier is weighted according to its performance and combined with other weak classifiers using voting (for classification) or averaging (for regression) to create a final model. The main advantages of boosting are that it can use any classification algorithm as a base learner, reduce model instability and have high predictive performance. For this, the boosting algorithm, based on a multiplicative weight-update technique [30], has been successfully applied to several benchmark machine learning problems using supervised learning. Basically, a minimization algorithm such as the least square (LS) can be used to boost for a strong learner from combining multiple weak learners whereby a new classifier is created based on the result of the previously generated classifiers by focusing on misclassified samples. The algorithm increases the weights of incorrectly classified samples and decreases the weights of those classified correctly. The LS boosting problem can be formulated as follows. Let x denote the feature vector and y the alignment accuracy. Given an input variable x, a response variable y and some samples $\{y_i, x_i\}_{i=1}^{N}$, the goal is to obtain an estimate or approximation $\hat{F}(x)$, of the function $F^*(x)$ mapping x to y, that minimizes some specified loss function ($L(y, F(x))$ over the joint distribution of all (y, x) values.

$$F^* = \arg\min_{F} L(y, F(x)), \tag{11}$$

where the squared error loss is given by $L(y, F) = (y-F)^2/2$ and the pseudo-response is obtained as

$$\tilde{y} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} = y_i - F_{m-1}(x), \quad i = 1, 2, ..., N. \tag{12}$$

Thus, for $i = 1, 2, ..., N$ the minimization of the data based estimate of the expected loss gives

$$(\rho_m, a_m) = \arg\min_{a,\rho} \sum_{i=1}^{N} [\tilde{y}_i - \rho h(x_i; a)]^2, \tag{13}$$

where $h(x; a)$ is the weak learner with basis functions $\{h(x, a_m)\}_{m=1}^{M}$ and ρ_m is the corresponding multiplier. The LS-boost algorithm [31], tuned to the problem of vehicular emissions prediction, has been described in [11], using Boosting Multivariate Adaptive Regression Splines (BMARS).

2.4. CART-BMARS hybrid modelling

Here, we propose to incorporate a regression tree with CART modelling to the BMARS algorithm [11] for improving the performance of air pollution prediction. CART builds the regression trees for predicting continuous dependent variables in the regression model. In this hybrid technique, the data sets are first passed through CART to generate node information. The training data together with node information are then supplied for training the BMARS. A rationale for the integration of CART and BMARS is that from a practical point of view, CART has the ability to handle missing values in the database by substituting surrogate splitters which are back-up rules to closely mimic the action of the primary splitting rule. This feature is not shared by many artificial intelligence approaches [12]. A flowchart of the proposed CART-BMARS hybrid model is shown in Fig. 3, wherein boosting is adopted to improve estimation performance by adjusting the weights of the classifiers.

2.5. ANN modelling

In order to verify merits of the CART-BMARS hybrid model, an ANN-based model is constructed to compare their predictive capabilities. In the present study, the multilayer feedforward neural network is trained by the back-propagation network (BPN) algorithm to correctly classify the training pair. Here, the Levenberg-Marquardt algorithm with a log-sigmoid activation function is used to update the network weights due to its high generalization capability. It is important to determine the optimum network architecture to achieve reliable results. This task still relies on trial-and-error even though several heuristic relations have been proposed to determine appropriately

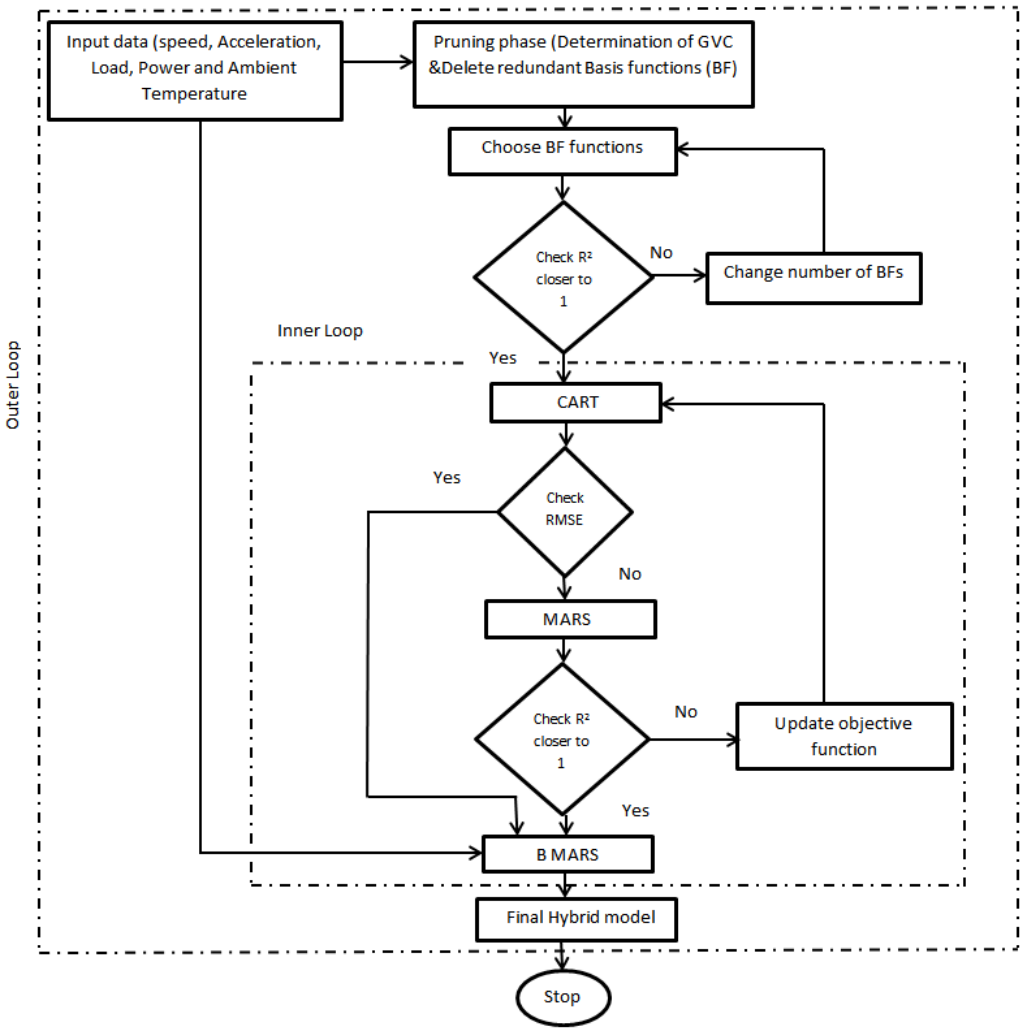


Figure 3: Flowchart of CART-BMARS hybrid model.

the number of neurons to be included in the hidden layer [32]. The architecture of the proposed ANNs is presented in Fig. 4, wherein the inputs are vehicles' speed, acceleration, load, power, and ambient temperature, and the outputs include NO_X, CO, CO_2 and THC. Here, the Root Mean Square Error (RMSE) is chosen as the loss function to be minimized, as RMSE possesses properties of convexity, symmetry, and differentiability for an excellent metric in the context of optimization.

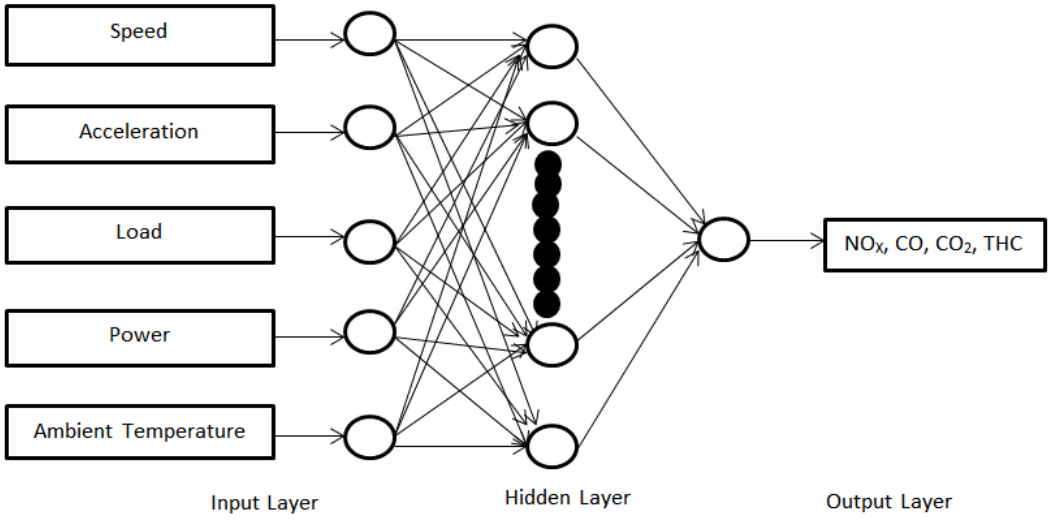


Figure 4: Proposed ANN architecture

3. Vehicular emission information and statistical evaluation

This section presents the collection of vehicular emissions data and preparation of datasets as well as the statistical evaluation of the output parameters.

3.1. Data collection

Vehicular emissions data used in this study were supplied by the Road and Maritime Service (RMS) of the New South Wales (NSW) Department of Vehicle Emission, Compliance Technology Operation. Ten (10) vehicles were used for the test, whereby emissions data were collected on the second by second basis. The test vehicles include Toyota, Mitsubishi, Holden, Ford and Nissan from 2009 and 2010 models with an engine displacement ranging from 1.8 L to 2.0 L. Emissions from these vehicles were recorded in two ways, by using a chassis dynamometer set-up and using a Horiba On-Board Measurement System (OBS-2000). Each drive cycle lasted for 556 seconds with the corresponding measurement of 556 data points.

The laboratorial dynamometer set-up was coupled to drive lines connected directly to the wheel hubs of the vehicle via a set of rollers upon which the vehicle was placed. These rollers can be adjusted to simulate driving resistance. During testing, the vehicle was tied down so that it remained stationary as a driver operated it according to a predetermined time-speed

profile for a given gear change pattern displayed on a monitor. The vehicle was considered as being driven to match the speed required at different stages of the driving cycle since experienced drivers are able to closely match an established speed profile.

The same vehicles were also tested with the Horiba On-Board Measurement System (OBS-2000). The equipment was composed of two on-board gas analysers, a laptop computer equipped with a data logger software, a power supply unit, a tailpipe attachment, and other accessories. The OBS-2000 collected the emission data via a global positioning system (GPS). Although the instrument measured other air pollutants, the focus of this paper was on such gases as CO, CO_2, THC and NO_X emissions. For logging the correct values of the measured emissions and other required parameters, the software was configured to a set of values provided by the Horiba Instruments, Inc. In addition, a delay in the logging attributed to the time it took to convert the measured concentrations from the analog to digital output was also accounted for by Horiba with appropriate adjustments in the data analysis spreadsheets.

3.2. Preparation of training datasets

The same datasets used for CART-BMARS hybrid model analysis are applied here for modelling and evaluating the prediction performance of the ANN-based model. Training a neural network architecture can be seen as a nonlinear optimization problem in which the task is to find out the set of parameters, i.e. synaptic weights, such that the network output is as close as to the desired output. Notably, the 556 values in the experimental dataset obtained were subject to a secondary emission correction by NSW RMS. Previous studies have shown that different ratios for training and testing data were required [33]. In the present study, 70% (390) of total experimental data was randomly selected for training the neural network, 15% (83) for the network cross-validation to avoid over-fitting, and the remaining 15% (83) of the data for testing the performance of the trained network. The data were first normalized as

$$R_N = \frac{R_A - R_{min}}{R_{max} - R_{min}}, \tag{14}$$

where R_A is the actual value, R_{min} and R_{max} are the minimum and maximum values of R, and R_N is the normalized value of R obtained within the range from 0 to 1.

3.3. Statistical evaluation of output parameters

After normalization, data were randomized and the ANN was trained and tested against the experimental data of vehicular emissions. In order to evaluate the prediction performance of the proposed ANN model, we have considered the R-squared, or correlation coefficient of determination (R^2) as a validation criterion:

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{N} (t_i - y_i)^2}{\sum_{i=1}^{N} (y_i)^2} \right). \tag{15}$$

The performance of the ANN-based predictions is evaluated by regression analysis of the predicted outputs and the target outputs. The correlation coefficient R^2 is used to assess the strength of this relationship, of which values closer to + 1 indicate a stronger positive linear relationship. Discrepancies between the predicted outputs (y) and the target outputs (t) are judged by the root mean squared error (RMSE):

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - t_i)^2}, \tag{16}$$

where N is the number of the data used for validation, t is actual output and y is the predicted output value.

4. Results and discussions

Based on the experimental data from the on-board system (OBS) and dynamometer (DYN) testing, the proposed CART-BMARS hybrid model is implemented for emissions prediction, where speed, acceleration, power, load and ambient temperature are used as predictors with different air pollutant emissions such as NO_X, CO, CO_2 and THC emissions. The results of the hybrid model computed using all the available data for on-board and dynamometer testing appear to have similar interpretations. It can be observed that all the five predictor variables play crucial roles in predicting the vehicle emissions by using all mentioned models. However, an analysis of variance (ANOVA) from the MARS model indicated that the two most important

variables were load and speed with acceleration, power and ambient temperature having less effects to emissions. To ensure a fair comparison, each time, the same training and test datasets were used for each model. The LS-boost algorithm for regressions with squared error loss is implemented in this paper. In the following, predictive performance of the models are compared in two perspectives. First, three learning techniques, namely CART, MARS and BMARS, are compared with the hybrid one to examine the best performance for emissions prediction. Then, a comprehensive analysis is conducted to demonstrate the predictive performance of the proposed CART-BMARS model against the ANNs one.

4.1. Comparison of CART-BMARS with BMARS, MARS and CART

Table 1 and Table 2 list respectively the RMSE and R-squared of the proposed CART-BMARS hybrid model in comparison with CART, MARS and BMARS models in terms of such pollutants as NO_X, CO, CO_2 and THC for both on-board system (OBS) and dynamometer testing (DYN). By analyzing the results in these tables, we can see that hybrid and BMARS models have smaller RMSE as compared to CART and MARS ones. Combining CART and MARS with boosting techniques as the hybrid model makes the algorithm relatively insensitive to the number of iterations, and their R^2 and RMSE values remain within a relatively stable range. Because both algorithms are forward additive, they can adaptively search for optimal results during the modelling process; this makes the model stable throughout the iteration range. Here, boosting turns the weaker classifier in the emission predicted variables into stronger classifier [19] and then builds many complement classifiers in order to find a highly accurate classifier on the training

Table 1: Comparison of HYBRID, BMARS, MARS and CART model (RMSE).

Model	HYBRID (RMSE)	BMARS (RMSE)	MARS (RMSE)	CART (RMSE)
NO_X-OBS	1.001×10^{-4}	3.367×10^{-4}	4.243×10^{-4}	4.244×10^{-4}
NO_X-DYN	2.478×10^{-4}	3.411×10^{-4}	4.652×10^{-4}	5.276×10^{-4}
CO-OBS	1.041×10^{-4}	2.945×10^{-4}	3.872×10^{-4}	4.145×10^{-4}
CO-DYN	2.143×10^{-4}	3.254×10^{-4}	5.276×10^{-4}	6.243×10^{-4}
CO_2-OBS	1.214×10^{-4}	2.845×10^{-4}	3.992×10^{-4}	4.249×10^{-4}
CO_2-DYN	2.478×10^{-4}	3.214×10^{-4}	4.652×10^{-4}	4.356×10^{-4}
THC-OBS	1.015×10^{-4}	2.946×10^{-4}	3.978×10^{-4}	3.284×10^{-4}
THC-DYN	2.784×10^{-4}	4.002×10^{-4}	5.115×10^{-4}	5.013×10^{-4}

Table 2: Comparison of HYBRID, BMARS, MARS and CART model (R^2).

Model	HYBRID (R^2)	BMARS (R^2)	MARS (R^2)	CART (R^2)
NO_X-OBS	0.951	0.739	0.624	0.624
NO_X-DYN	0.818	0.706	0.593	0.504
CO-OBS	0.962	0.757	0.656	0.608
CO-DYN	0.881	0.723	0.504	0.476
CO_2-OBS	0.906	0.774	0.642	0.623
CO_2-DYN	0.854	0.656	0.608	0.608
THC-OBS	0.907	0.757	0.672	0.656
THC-DYN	0.809	0.608	0.518	0.534

set by ensembling the weak hypotheses. The outcome of the proposed model is a higher R^2 value and lower RMSE. The selection of the generalized cross-validation GCV criterion in both models tends to be sensitive and can overfit the model. Obviously, these relatively poor selections will degrade the model results in CART or MARS models, so the results are unstable with a lower accuracy [29]. This suggests the robustness of the hybrid algorithm and its capability of improving accuracy of the MARS model in vehicular emissions prediction.

In general, boosting can improve the prediction accuracy of a particular learning model. As can be seen, the performance of CART-BMARS hybrid and BMARS is better than that of CART or MARS alone. Comprehensive analysis shows that, combining CART with BMARS to form a hybrid model is superior to a non-boosting strategy, or a strategy without a regression tree. Furthermore, this proves that the hybrid model strategy used in this paper has effectively improved the prediction accuracy and generalization ability of the emission models. From Tables 1 and 2, it can be observed that the hybrid model for all pollutants NO_X, CO, CO_2 and THC using both DYN and OBS systems outperformed the BMARS in terms of goodness of fit and prediction accuracy. The hybrid model also takes the advantage of BMARS in its capability of handling non-linearity in the data. However, since boosting is sensitive to noise within data, its performance may be affected in the presence of noise, depending on the boosting method used. While the MARS have a potential problem of over-fitting the model and hence, subject to computational complexity, the CART-BMARS hybrid model can effectively handle the corresponding noise interval and the missing values

within the database at every step. Thus, the proposed technique is able to adequately solve the problem of boosting's sensitivity to data noise and ultimately improve the prediction performance of the emissions model. This explains why its prediction performance is better than BMARS.

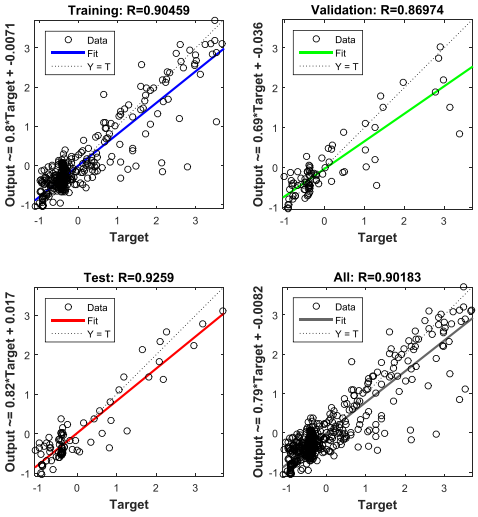
4.2. Comparison of CART-BMARS with ANNs

In this work, we use the BPN which is adequate for predicting vehicular emissions. The accuracy of neural network prediction is generally dependent on the number of hidden layers and the numbers of neurons in each layer. To find out the suitable architecture, a number of neural network architectures have been tested by varying the number of hidden neurons from 2 to 15 with 5 inputs (speed, acceleration, load, power and ambient temperature) and 1 output respectively for each pollutant (NO_X, CO, CO_2 and THC).

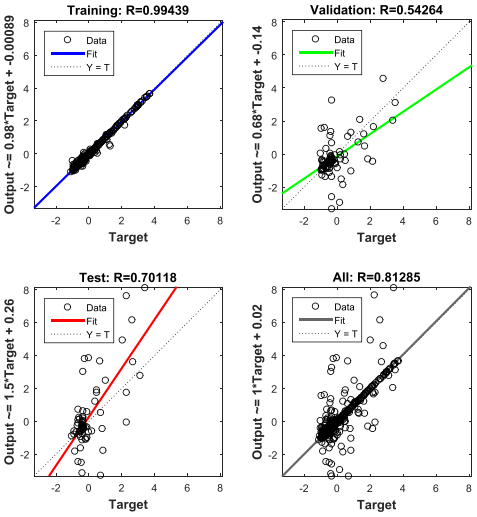
Table 3: ANNs architecture and prediction accuracy.

Model	Hidden layer neuron number	RMSE
ANN-NO_X-OBS	11	2.244×10^{-4}
ANN-NO_X-DYN	12	3.921×10^{-4}
ANN-CO-OBS	8	2.278×10^{-4}
ANN-CO-DYN	10	3.651×10^{-4}
ANN-CO_2-OBS	9	2.375×10^{-4}
ANN-CO_2-DYN	13	2.952×10^{-4}
ANN-THC-OBS	7	2.662×10^{-4}
ANN-THC-DYN	14	2.978×10^{-4}

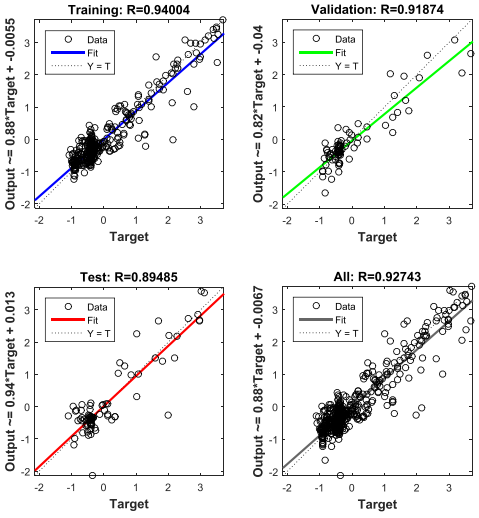
The results are listed in Table 3, showing the air pollutant models, the number of hidden neurons correspondingly, and the accuracy in terms of RMSE. The remaining data, set aside for testing and validation purposes, were then used to check the predictive capabilities of the trained model. Comparison of the output obtained by the ANNs and the target values of the experimental data are shown in regression plots of Figs. 5 and 6 for all four air pollutants with on-board and dynamometer testing systems. As observed from the graphs, a high correlation between the predicted and the experimental values demonstrates that the model succeeded in predicting major emissions from vehicles. The regression plots yield high R^2 values closer to 1 for both on-board and dynamometer tests, indicating that ANNs are a useful method for prediction of vehicular emissions.



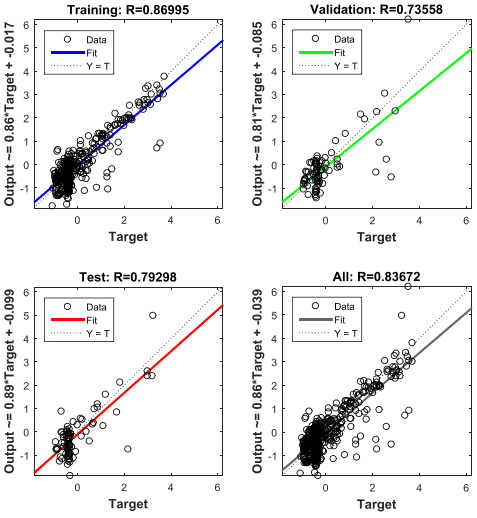
(a) NO_X on-board



(b) NO_X dynamometer

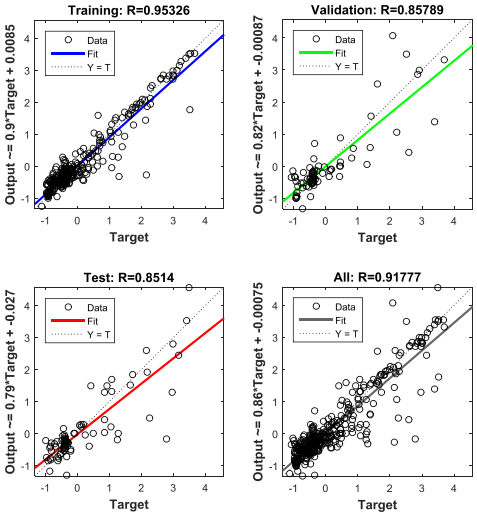


(c) CO on-board

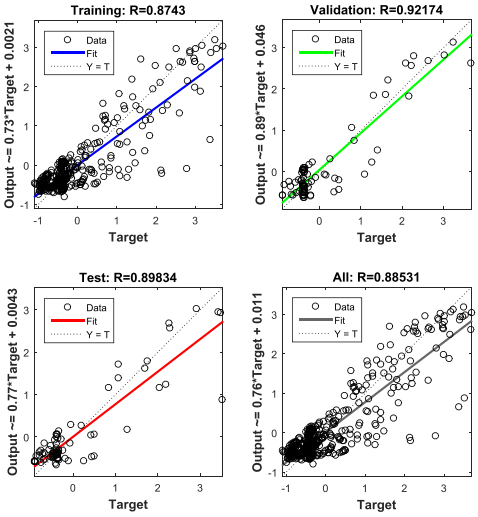


(d) CO dynamometer

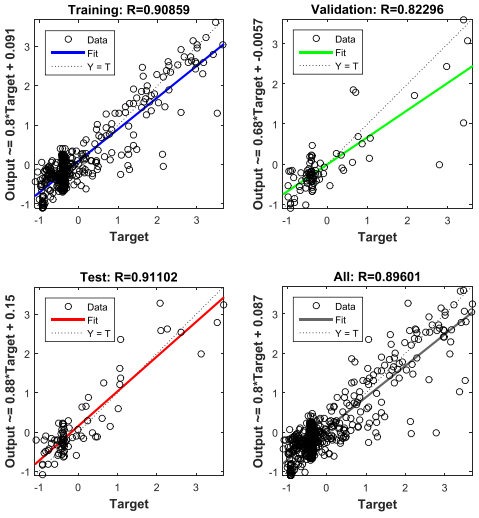
Figure 5: Regression plots corresponding to the designed ANNs model.



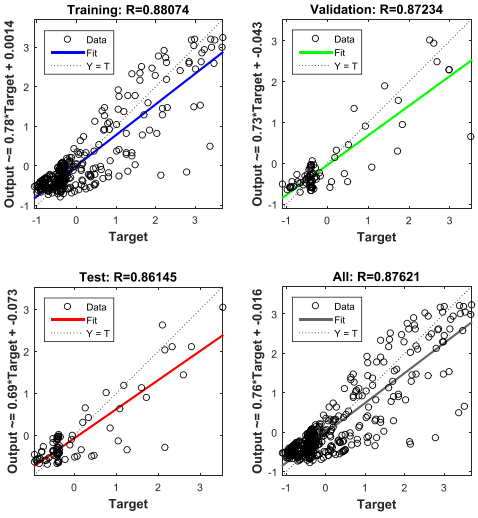
(a) CO_2 on-board



(b) CO_2 dynamometer



(c) THC on-board



(d) THC dyanmometer

Figure 6: Regression plots corresponding to the designed ANNs model

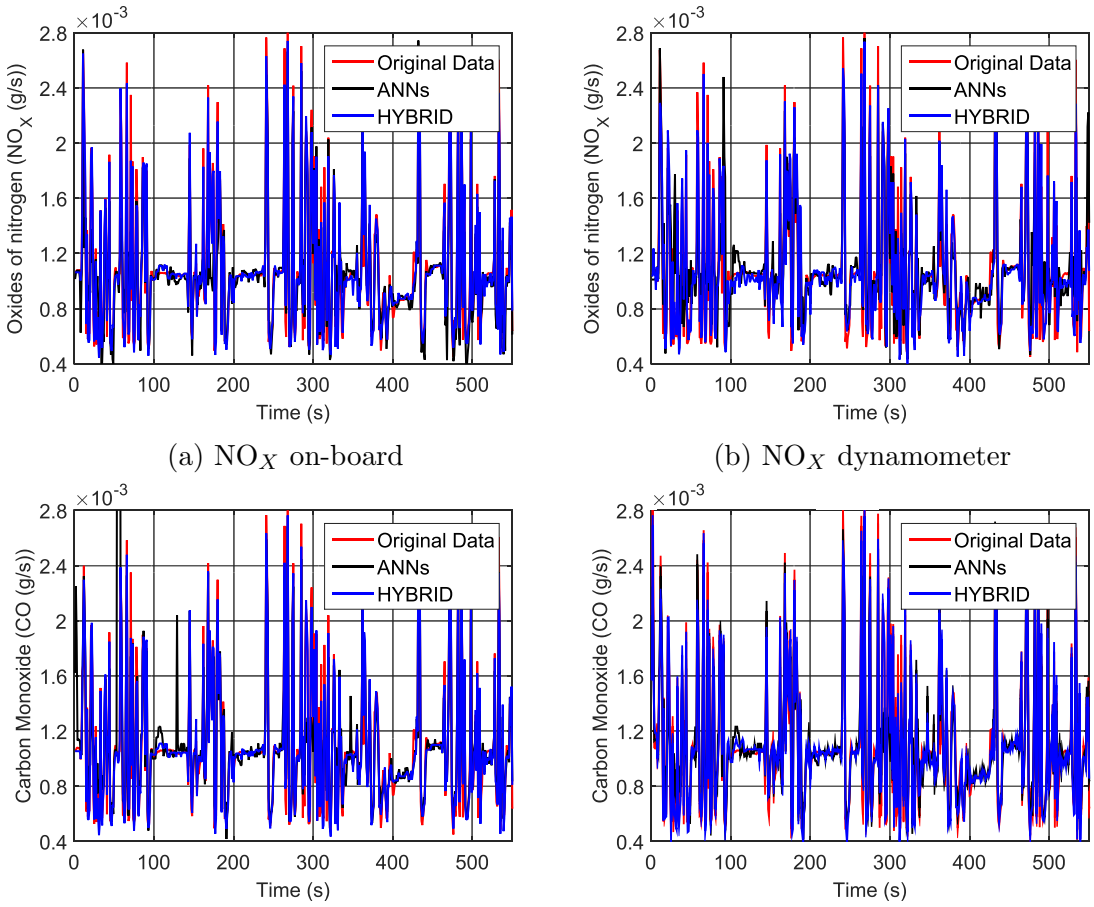
The performance of the proposed CART-BMARS and ANNs models were compared with the experimental dataset, as shown in Figs. 7 and 8, respectively for NO_X, CO, CO_2 and THC. As observed, the results obtained show

Table 4: Performance Comparison between Hybrid and ANNs models

Model	Processing Time (s)	R^2
Hybrid-NO_X-OBS	6	0.951
Hybrid-NO_X-DYN	8	0.817
ANN-NO_X-OBS	22	0.814
ANN-NO_X-DYN	23	0.659
Hybrid-CO-OBS	7	0.962
Hybrid-CO-DYN	9	0.879
ANN-CO-OBS	19	0.859
ANN-CO-DYN	20	0.701
Hybrid-CO_2-OBS	7	0.904
Hybrid-CO_2-DYN	9	0.854
ANN-CO_2-OBS	21	0.843
ANN-CO_2-DYN	24	0.783
Hybrid-THC-OBS	7	0.906
Hybrid-THC-DYN	10	0.808
ANN-THC-OBS	23	0.803
ANN-THC-DYN	25	0.767

excellent performance indices for both CART-BMARS hybrid and ANNs models and are also in agreement with other researchers using the same methodology for different applications ([34], [35]).

The models' performance and the efficiency features are listed in Table 4 for comparison of CART-MARS and ANNs merits. The results therein together with those in Table 1 and Table 3 confirm the advantages of the CART-BMARS hybrid model method for all pollutant emissions considered. In addition, it appears to be faster than ANNs as the processing speed (CPU time) remains smaller for all cases, as shown in Table 4. Another distinctive aspect is that it can identify the contribution of each variable to the emissions prediction through the analysis of variance (ANOVA) decomposition. The model output is expressed in a more interpretable way in the form of "segmented" defined on different intervals and may provide additional information about how changes in the input data can affect the output.



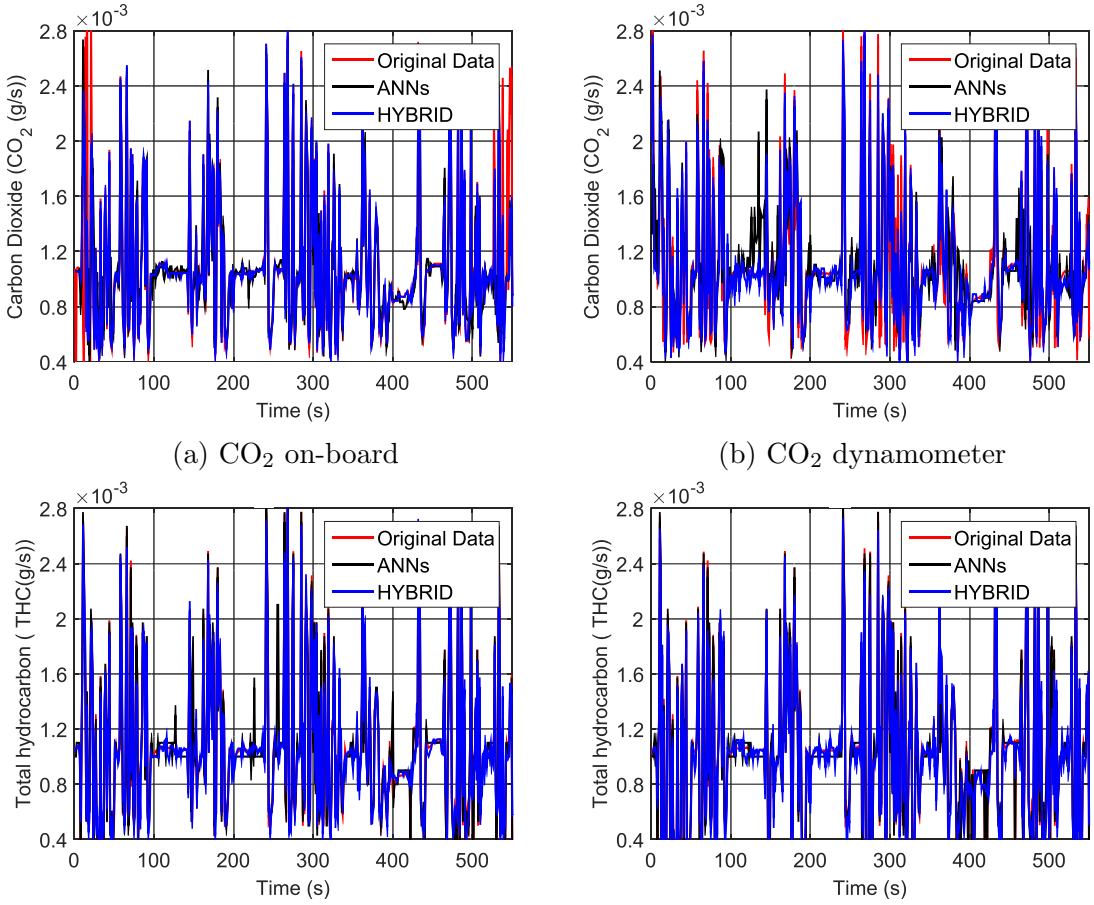
(a) NO_X on-board

(b) NO_X dynamometer

(c) CO on-board

(d) CO dynamometer

Figure 7: Comparison of Hybrid and ANNs models with experimental data for NO_X and CO emissions.



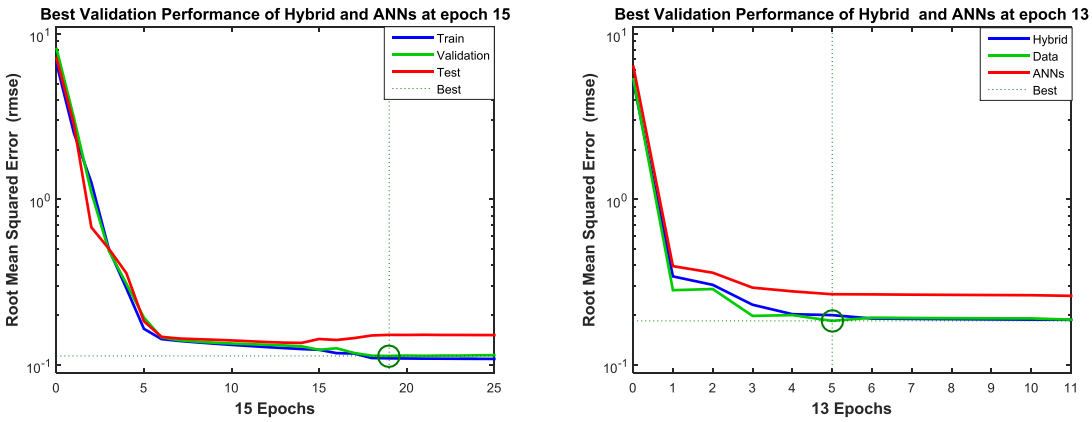
(a) CO_2 on-board

(b) CO_2 dynamometer

(c) THC on-board

(d) THC dynamometer

Figure 8: Comparison of Hybrid and ANNs with experimental data for CO_2 and THC emissions.



(a) On-board (b) Dynamometer

Figure 9: Performance evaluation of Hybrid and ANN models.

The effectiveness of the combination of CART and BMARS as a hybrid method for vehicular emissions can be explained as (i) the proposed hybrid model is computationally more efficient owing to the capability of dividing the predictors space into multiple knots and then fitting a spline function between them, and hence, requiring less trial and error as compared to the ANNs model, and (ii) the CART-BMARS technique allows for effectively removing data noise and reducing the sensitivity of boosting to noise in the emissions data, while the final number of basis functions can be determined form a preset maximum value. The performance of the proposed method in comparison against the ANNs model is further evaluated in terms of RMSE, as shown in Fig. 9. Therein, the hybrid model RMSE (in blue) reduces gradually and gets closer to the validation data (in green) unlike that of the ANNs model (in red) for both the on-board and dynamometer testing systems. These results suggest that the CART-BMARS hybrid model can constitute a valuable alternative for predicting vehicular emissions.

5. Conclusion

In this paper, a CART-BMARS hybrid method has been proposed to estimate the nonlinear relationship between vehicular pollutant emissions and predictor variables such as speed, acceleration, load, power and ambient temperature as predictor variables. The hybrid model is implemented with effective piecewise-linear BFs which effectively solve the problem of non-linearity

and uncertainty in the emissions data and improve the prediction accuracy of the model. The new hybrid model is developed to overcome the shortcomings of MARS and BMARS models, effectively improving the performance the emissions model. The proposed hybrid algorithm is then compared with a multilayer BPN trained and tested by the Levenberg-Marquardt optimization algorithm. It can be observed that among all techniques mentioned, the proposed CART-BMARS hybrid model exhibits excellent prediction performance for all pollutant emissions using both on-board and dynamometer testing systems.

Acknowledgements

The data used for this study were provided by the Road and Maritime Service, Department of Vehicle Emission, Compliance Technology & Compliance Operations, NSW Office of Environment & Heritage, and HORIBA, Australia.

References

[1] M. Azzi, H. Duc, and Q. Ha, "Towards sustainable energy usage in the power generation and construction sectors - a case study of australia," *Automation in Construction*, vol. 59, pp. 122–127, 2015.

[2] IEA, *World energy outlook (2008)*. Paris: International Energy Agency, 2009.

[3] A. Elkafoury, A. M. Negm, M. F. Bady, and M. H. Aly, "Modeling vehicular CO emissions for time headway-based environmental traffic management system," *Procedia Technology*, vol. 19, pp. 341 – 348, 2015.

[4] Y. Tong, X. Wang, J. Zhai, X. Niu, and L. Liu, "Theoretical predictions and field measurements for potential natural ventilation in urban vehicular tunnels with roof openings," *Building and Environment*, vol. 82, pp. 450 – 458, 2014.

[5] N. V. Sonawane, R. S. Patil, and V. Sethi, "Health benefit modelling and optimization of vehicular pollution control strategies," *Atmospheric Environment*, vol. 60, pp. 193 – 201, 2012.

[6] R. Goel and S. K. Guttikunda, "Evolution of on-road vehicle exhaust emissions in delhi," *Atmospheric Environment*, vol. 105, pp. 78 – 90, 2015.

[7] N. Holmes and L. Morawska, "A review of dispersion modelling and its application to the dispersion of particles: An overview of different dispersion models available," *Atmospheric Environment*, vol. 40, no. 30, pp. 5902 – 5928, 2006.

[8] H. C. Frey and K. Kim, "In-use measurement of the activity, fuel use, and emissions of eight cement mixer trucks operated on each of petroleum diesel and soy-based b20 biodiesel," *Transportation Research Part D: Transport and Environment*, vol. 14, no. 8, pp. 585 – 592, 2009.

[9] S. Pandian, S. Gokhale, and A. K. Ghoshal, "Evaluating effects of traffic and vehicle characteristics on vehicular emissions near traffic intersections," *Transportation Research Part D: Transport and Environment*, vol. 14, no. 3, pp. 180 – 196, 2009.

[10] B. Y. Boroujeni and H. C. Frey, "Road grade quantification based on global positioning system data obtained from real-world vehicle fuel use and emissions measurements," *Atmospheric Environment*, vol. 85, pp. 179 – 186, 2014.

[11] S. Oduro, S. Metia, H. Duc, G. Hong, and Q. Ha, "Multivariate adaptive regression splines models for vehicular emission prediction," *Visualization in Engineering*, vol. 3, no. 1, 2015.

[12] H. Li, J. Sun, and J. Wu, "Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods," *Expert Systems with Applications*, vol. 37, no. 8, pp. 5895 – 5904, 2010.

[13] M. Elbayoumi, N. A. Ramli, and N. F. F. M. Yusof, "Development and comparison of regression models and feedforward backpropagation neural network models to predict seasonal indoor PM2.5 concentrations in naturally ventilated schools," *Atmospheric Pollution Research*, vol. 6, no. 6, pp. 1013 – 1023, 2015.

[14] S. S. Nagendra and M. Khare, "Artificial neural network approach for modelling nitrogen dioxide dispersion from vehicular exhaust emissions," *Ecological Modelling*, vol. 190, no. 12, pp. 99 – 115, 2006.

[15] G. Najafi, B. Ghobadian, T. Tavakoli, D. Buttsworth, T. Yusaf, and M. Faizollahnejad, "Performance and exhaust emissions of a gasoline engine with ethanol blended gasoline fuels using artificial neural network," *Applied Energy*, vol. 86, no. 5, pp. 630 – 639, 2009.

[16] B. Ghobadian, H. Rahimi, A. Nikbakht, G. Najafi, and T. Yusaf, "Diesel engine performance and exhaust emission analysis using waste cooking biodiesel fuel with an artificial neural network," *Renewable Energy*, vol. 34, no. 4, pp. 976 – 982, 2009.

[17] A. Kurt and A. B. Oktay, "Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks," *Expert Systems with Applications*, vol. 37, no. 12, pp. 7986 – 7992, 2010.

[18] Q. Ha, H. Wahid, H. Duc, and M. Azzi, "Enhanced radial basis function neural networks for ozone level estimation," *Neurocomputing*, vol. 155, pp. 62–70, 2015.

[19] L. Breiman, H. L. Friedmann, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Wadsworth International Group, 1984.

[20] B. W. Yap, S. H. Ong, and N. H. M. Husain, "Using data mining to improve assessment of credit worthiness via credit scoring models," *Expert Systems with Applications*, vol. 38, no. 10, pp. 13 274 – 13 283, 2011.

[21] M.-Y. Chen, "Predicting corporate financial distress based on integration of decision tree classification and logistic regression," *Expert Systems with Applications*, vol. 38, no. 9, pp. 11 261 – 11 272, 2011.

[22] A. Tayyebi and B. C. Pijanowski, "Modeling multiple land use changes using ANN, CART and MARS: Comparing tradeoffs in goodness of fit

and explanatory power of data mining tools," *International Journal of Applied Earth Observation and Geoinformation*, vol. 28, pp. 102 – 116, 2014.

[23] M. K. Ayoubloo, H. M. Azamathulla, E. Jabbari, and M. Zanganeh, "Predictive model-based for the critical submergence of horizontal intakes in open channel flows with different clearance bottoms using CART, ANN and linear regression approaches," *Expert Systems with Applications*, vol. 38, no. 8, pp. 10 114 – 10 123, 2011.

[24] W.-Y. Loh, "Tree-structured classifiers," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 3, pp. 364–369, 2010.

[25] P. Paulsen, F. Smulders, A. Tichy, A. Aydin, and C. Hck, "Application of classification and regression tree CART analysis on the microflora of minced meat for classification according to reg. EC 2073/2005," *Meat Science*, vol. 88, no. 3, pp. 531 – 534, 2011.

[26] J. H. Friedman, "Multivariate adaptive regression splines," *Annals of Statistics*, vol. 19, no. 1, pp. 1–144, 1991.

[27] M. Abdel-Ati and K. Haleem, "Analyzing angle crashes at unsignalized intersections using machine learning techniques," *Accident Analysis and Prevention*, vol. 43, pp. 461–470, 2011.

[28] R. Put, Q. Xu, D. Massart, and Y. V. Heyden, "Multivariate adaptive regression splines MARS in chromatographic quantitative structure-retention relationship studies," *Journal of Chromatography A*, vol. 1055, no. 12, pp. 11 – 19, 2004.

[29] R. Hastie, T. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer-Verlag, 2009.

[30] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, pp. 119–139, 1997.

[31] H. F. Jerome, *Greedy Function Approximation: A Gradient Boosting Machine*. Institute of Mathematical Statistics: Prentice-Hall, 2001.

[32] H. Rafiai and M. Moosavi, "An approximate ann-based solution for convergence of lined circular tunnels in elasto-plastic rock masses with anisotropic stresses," *Tunnelling and Underground Space Technology*, vol. 27, no. 1, pp. 52 – 59, 2012.

[33] S. Oduro, S. Metia, H. Duc, and Q. Ha, "Predicting carbon monoxide emissions with multivariate adaptive regression splines MARS and artificial neural networks ANNs," in *The 32nd International Symposium on Automation and Robotics in Construction and Mining*, June 2015, pp. 912–920.

[34] C. Manzie, H. Watson, and S. Halgamuge, "Fuel economy improvements for urban driving: Hybrid vs. intelligent vehicles," *Transportation Research Part C: Emerging Technologies*, vol. 15, no. 1, pp. 1 – 16, 2007.

[35] M. Sorek-Hamer, A. Strawa, R. Chatfield, R. Esswein, A. Cohen, and D. Broday, "Improved retrieval of PM2.5 from satellite data products using non-linear methods," *Environmental Pollution*, vol. 182, pp. 417 – 423, 2013.