

Identification of the Dynamic Operating Envelope of HCCI Engines Using Class Imbalance Learning

Vijay Manikandan Janakiraman, XuanLong Nguyen, Jeff Sterniak, and Dennis Assanis

Abstract—Homogeneous charge compression ignition (HCCI) is a futuristic automotive engine technology that can significantly improve fuel economy and reduce emissions. HCCI engine operation is constrained by combustion instabilities, such as knock, ringing, misfires, high-variability combustion, and so on, and it becomes important to identify the operating envelope defined by these constraints for use in engine diagnostics and controller design. HCCI combustion is dominated by complex nonlinear dynamics, and a first-principle-based dynamic modeling of the operating envelope becomes intractable. In this paper, a machine learning approach is presented to identify the stable operating envelope of HCCI combustion, by learning directly from the experimental data. Stability is defined using thresholds on combustion features obtained from engine in-cylinder pressure measurements. This paper considers instabilities arising from engine misfire and high-variability combustion. A gasoline HCCI engine is used for generating stable and unstable data observations. Owing to an imbalance in class proportions in the data set, the models are developed both based on resampling the data set (by undersampling and oversampling) and based on a cost-sensitive learning method (by overweighting the minority class relative to the majority class observations). Support vector machines (SVMs) and recently developed extreme learning machines (ELM) are utilized for developing dynamic classifiers. The results compared against linear classification methods show that cost-sensitive nonlinear ELM and SVM classification algorithms are well suited for the problem. However, the SVM envelope model requires about 80% more parameters for an accuracy improvement of 3% compared with the ELM envelope model indicating that ELM models may be computationally suitable for the engine application. The proposed modeling approach shows that HCCI engine misfires and high-variability combustion can be predicted ahead of time, given the present values of available sensor measurements, making the models suitable for engine diagnostics and control applications.

Manuscript received August 9, 2013; revised February 27, 2014; accepted March 6, 2014. Date of publication April 1, 2014; date of current version December 16, 2014. This work was supported by the Department of Energy and performed as a part of the ACCESS project consortium (Robert Bosch LLC, AVL Inc., Emitec Inc.) under the direction of PI Hakan Yilmaz, Robert Bosch, LLC. The work of X. Nguyen was supported by NSF under Grant CCF-1115769 and Grant ACI-1047871.

V. M. Janakiraman was with the Department of Mechanical Engineering, University of Michigan, Ann Arbor, MI 48109 USA. He is now with UARC, NASA Ames Research Center, Moffett Field, CA 94040 USA (e-mail: vijai@umich.edu).

X. Nguyen is with the Department of Statistics, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: xuanlong@umich.edu).

J. Sterniak is with Robert Bosch LLC, Farmington Hills, MI 48331 USA (e-mail: jeff.sterniak@us.bosch.com).

D. Assanis is with Stony Brook University, NY 11794 USA (e-mail: dennis.assanis@stonybrook.edu).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TNNLS.2014.2311466

Index Terms—Class imbalance learning, dynamic classification, engine control, engine diagnostics, extreme learning machine, homogeneous charge compression ignition, misfire prediction, operating envelope model, support vector machine, system identification.

I. INTRODUCTION

HOMOGENEOUS charge compression ignition (HCCI) engines are of significant interest to the automotive industry owing to their ability to reduce emissions and fuel consumption significantly compared with the traditional spark ignition and compression ignition engines [1]–[3]. The highly efficient operation of HCCI is achieved using advanced control strategies, such as exhaust gas recirculation [4], variable valve timings (VVTs) [5], and intake charge heating [6] among others. Such complex manipulations of the system result in a highly nonlinear behavior [7] and a narrow region of stable operation [8], [9].

The HCCI engine is constrained to operate between operating limits, such as misfire, ringing, knock, and so on [10], [11]. In order to develop controllers and operate the engine in a stable manner, it is imperative that the admissible operating envelope of the system be determined. The operating envelope of an engine can be defined as a region in the input space (system actuators and sensed physical variables of the engine) that results in a stable operation of the engine. Knowledge of the operating envelope is crucial for designing efficient controllers for the following reasons. The developer can get insights on the actuator extremes [12], for instance, the minimum and maximum quantity of fuel to be injected into the engine at a given speed and load conditions. Such information can be used to enforce constraints on the control variables for desired engine operation. In addition, the operating envelope model can act as a filter to perform system identification by eliminating excitations that might lead the system to be unstable. HCCI engine is a partially stable nonlinear system. For system identification, the system needs to be excited at several amplitudes and frequencies. It has been observed previously by the authors that HCCI engine leads to misfires or high-variability combustion modes when excited at undesirable input combinations [13], [14]. A model of the admissible operating envelope of the engine can help to design input excitations for system identification of the HCCI engine. Furthermore, an operating envelope model could enable designing efficient engine diagnostic systems based on

predictive analytics. For instance, a misfire event is a lack of combustion that produces no work output from the engine. The misfired fuel enters the exhaust system increasing emissions of hydrocarbon and carbon monoxide [15], [16]. When the engine misfires, pollutant levels may be higher than normal. Real-time monitoring of the exhaust emission control system and engine misfire detection are essential to meet requirements on onboard diagnostic regulations. The proposed envelope model can be used to alarm the onboard diagnostics if the engine is about to misfire owing to changes in system or operating conditions.

HCCI operating envelope can be characterized using operating variables of the engine, such as air-to-fuel ratio, fuel injection timing, exhaust gas recirculation rate, fueling rate, and so on. The operating envelope of early simpler engines is characterized manually where the number of variables is less. To speed up the process, automated methods, such as limit search algorithms, were introduced [17]. However, with increasing engine complexity characterized by high-dimensional operating spaces, more sophisticated algorithms based on machine learning were introduced. The operating envelope of a gasoline direct injection engine was modeled using support vector machines (SVMs) [12]. The operating boundary was assumed to be a static map where time history was not considered. Such maps do not include engine transient effects, for instance, the high load limit of HCCI is significantly different when a dynamically changing variable, such as cylinder wall temperature, is either high or low [4]. The HCCI engine considered in this paper is achieved using an exhaust recompression strategy [13], i.e., the exhaust gases from the previous combustion cycles influence the future engine operation. As a result, a model that captures this dynamic behavior is required. To the best of the authors' knowledge, such a dynamic model of the operating envelope of the HCCI engine is unreported in the literature. As a matter of fact, owing to the complexity of combustion engines, a dynamic operating envelope of any engine is unavailable in the literature. This paper aims to present a novel approach based on statistical and predictive modeling of a dynamic operating envelope of combustion engines.

The problem considered in this paper is to develop a predictive model of the dynamic operating envelope of the HCCI engine. Specifically, the operating envelope defined by two common HCCI unstable modes—a complete misfire and a high-variability combustion (a more detailed description is given in Section III-B), is studied. The problem of identifying the HCCI operating envelope using the experimental data can be posed as a classification problem. The engine sensor data can be manually labeled as stable or unstable depending on engine-based heuristics. Furthermore, the engine dynamic data consist of a large number of stable class data compared with unstable class data, which introduce an imbalance in class proportions. As a result, the problem can be posed as a class imbalance learning (CIL) of a binary classification decision boundary.

Although classification algorithms were applied for a wide spectrum of applications, the literature on engine application involves only diagnostics-related problems [18]–[21]. An

SVM-based approach is the only reported work on identification of an engine operating envelope [12], where the envelope was assumed to be a static system, i.e., the measurement history was not considered resulting in a simple static binary classifier model. In contrast to misfire prediction [12], misfire detection algorithms based on fluctuations in engine speed [20], [21] are more common. The detectors detect misfires after the event has occurred as compared against the predictive models that predict misfires in future and can be used in onboard decision making and other predictive analysis. To this end, a novel solution is presented for modeling the dynamic operating envelope of combustion engines. This is the main contribution of this paper. A suitable modeling structure and framework to capture the nonlinear dynamical behavior of HCCI is identified purely based on data. A data labeling approach based on stability definitions is proposed based on combustion features calculated by available combustion analysis tools. Furthermore, in contrast to the literature where steady state data are used for operating envelope modeling, this paper deals with transient (dynamic) data from the engine, which is a significant extension. The application and adaptation of SVMs and recently developed extreme learning machines (ELMs) for developing dynamic classifiers based on imbalanced data sets for combustion engines is a novel application. The ELM algorithm is relatively new; moreover, very few engine applications are reported in [22] and [23]. This paper adds to this list by demonstrating ELM application including model selection, parameter tuning, and adaptation to the HCCI engine data. Finally, a working model that predicts the HCCI engine misfires and high-variability combustion in transient operation is an extension to the present state of the art that uses the steady-state information of the engine alone.

For the imbalance classifier design, the following two approaches have been evaluated:

- 1) heuristic resampling of data: apply preprocessing methods, such as undersampling and oversampling of data to get a balanced data set;
- 2) cost-sensitive approach: modify the objective function of the learning system to weigh the minority class data more heavily.

Four dynamic envelope models including SVMs, ELMs, logistic regression (LR), and linear least squares (LS) have been developed. The models are compared for generalization accuracy, along with storage required in the engine electronic control unit (ECU). The LS and LR envelope models were developed as baselines for comparison. This paper is organized as follows. A brief background on the classification algorithms along with cost-sensitive modifications is given in Section II. The HCCI engine experiments and data labeling are briefed in Section III with HCCI envelope modeling, prediction results, and model interpretation are discussed in Section IV followed by the conclusion in Section V.

II. CLASSIFICATION ALGORITHMS FOR HCCI DATA

In this section, a mathematical abstraction of the engine data is presented along with suitable modeling assumptions

for constructing learning models. The learning algorithms considered in this paper are introduced along with the necessary mathematical background.

In order to keep the discussion general, let the abstracted data from the engine be represented as $\{(x_1, y_1), \dots, (x_N, y_N)\} \in (\mathcal{X}, \mathcal{Y})$, where \mathcal{X} denotes the space of the input features (let $\mathcal{X} = \mathbb{R}^n$), while \mathcal{Y} represent class label values $\{-1, +1\}$, and N denotes the number of observations. The goal of the classification algorithm is to model the underlying boundary separating the data into two distinct classes by minimizing a risk function $R(w)$ with respect to the model parameters w

$$R(w) = \frac{1}{N} \sum_{i=1}^N L(y_i - \hat{y}_i(x|w)) + \frac{1}{2} w^T w. \quad (1)$$

Here, $R(w)$ has two components—the empirical risk representing the training error and the structural risk for the model parameters, L represents a loss function, and $\hat{y}(x|w, b)$ represents the model prediction, whose structures are given by the learning algorithm (see the following subsections). The algorithms considered in this paper are LR, SVMs, and ELMs. Each of the algorithms is unique in its formulation, loss function used, convergence rates, computational demand, prediction accuracy, and potential for online learning. However, the main criteria used for evaluation in this paper are prediction accuracy, number of parameters used for modeling, and potential for online adaptation onboard the engine ECU. The HCCI classification problem involves identifying the boundary separating the input space that results in a stable or unstable operation. In addition, when the engine misfires, the excitation command is changed to attempt a stable operation [14], where the engine is operated until the next command in the sequence is activated (refer to Section III-A for experiments). This results in a class imbalance data set as the number of unstable class data is significantly smaller than the number of stable class data.

A. Class Imbalance Learning

CIL is encountered during situations when the number of instances of one class is very different from the number of instances in others. In a binary classification problem, the class where the number of observations are large in number is referred to as the majority class (labeled $+1$), while the other class is referred to as the minority class (labeled -1). Imbalanced data sets need careful attention, as machine learning (typically an optimization problem) causes the decision boundary to be more biased toward the majority class data while ignoring the minority class data [24], [25].

Several solutions have been proposed to handle CIL problems, including resampling the data where the minority class can be duplicated to be in proportion with the majority class (referred to as oversampling), or some majority class data are removed to match proportions with the minority class (referred to as undersampling). Although both sampling methods aim to artificially obtain a balanced data set, undersampling is prone to loss of majority class information, while oversampling is prone to overfitting [24], [25]. Algorithm level modifications

are also common, which include cost-sensitive learning that weights the minority class data more than the majority class data in the optimization objective function. Other methods, such as adjusting the decision threshold, one-class learning and so on, are available in the literature, but the focus of this paper is the comparison of undersampling, oversampling and cost-sensitive methods.

B. Logistic Regression

LR is a classical linear classifier that proves to be effective especially for large data set problems owing to its computational efficiency. LR makes use of a logistic function given by (2), which confines the output of the function to lie between zero and one. Unlike the linear regression model that solves a LS problem with a squared loss function, LR solves a nonlinear optimization problem using a logistic loss function. The logistic loss function is particularly attractive for classification because the algorithm does not penalize the correctly classified points as much as the squared loss improving convergence

$$\psi(x) = \frac{1}{1 + e^{-x}}. \quad (2)$$

The conditional probability of estimating y from x can be expressed in terms of the model parameters $\beta = [\beta_0 \ \beta_1]^T$ as

$$P(Y = y|X = x) = \frac{1}{1 + e^{-y(\beta_1^T x + \beta_0)}} \quad (3)$$

where X and Y represent the input and output random variables. The goal of LR is to determine β such that $P(Y|X, \beta)$ is maximized using the following optimization problem:

$$\beta^* = \arg \min_{\beta} \sum_{i=1}^N \log \left(1 + e^{-y(\beta_1^T x + \beta_0)} \right). \quad (4)$$

Equation (4) is nonlinear in β and can be solved by simple iterative methods [26]. The LR decision hypothesis is given by

$$f(x) = \text{sgn}(\beta_1^T x + \beta_0) \quad (5)$$

where

$$\text{sgn}(x) = \begin{cases} 1 & x > 0 \\ -1 & x \leq 0. \end{cases} \quad (6)$$

C. Support Vector Machines

SVMs involve determining the boundary that maximizes the margin separating the classes of data. This is achieved by utilizing the so-called hinge loss $L_{\text{hinge}}(w, b) = \max(0, 1 - y f(x))$, where $y f(x)$ gives the margin [27]–[29]. This translates to finding the optimal model parameters (w^*, b^*) by solving the following optimization problem:

$$\min_{w, b, \zeta_i} \frac{1}{2} w^T w + C \sum_{i=1}^N \zeta_i \quad (7)$$

$$\text{s.t. } \begin{cases} y_i[(w, \phi(x_i)) + b] \geq 1 - \zeta_i \\ \zeta_i \geq 0 \end{cases} \quad (8)$$

for $i = 1, \dots, N$. Here, ζ_i is named a slack variable for data observation i , and C represents the cost penalty hyperparameter. The slack variables ζ_i are required in order to allow for misclassifications in a noisy overlapping binary data set that cannot be completely separated by a decision boundary. The input vectors x are mapped onto a higher dimensional space using the function ϕ . By making this transformation, the nonlinear data are aligned linearly in the high-dimensional space, where SVM finds a maximum margin separating hyperplane. The transformation is performed implicitly using a kernel matrix $K(x_i, x_j) = [k(x_i, x_j)]_{i,j}$, where $k(x_i, x_j)$ could be any function satisfying Mercer's condition [27]. The Gaussian kernel function (9) is used in this paper. More details on SVM formulation can be found in [27]

$$k(x_i, x_j) = e^{-\sigma \|x_i - x_j\|^2}, \sigma > 0. \quad (9)$$

The convex constrained optimization problem in (7) is in the primal form, and the variables w , b , and ζ_i are referred to as primal variables. The primal problem is converted to a dual formulation in (10) and solved for the dual variables α_i

$$\max_{\alpha_i} \left\{ -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N y_i y_j \alpha_i \alpha_j K(x_i, x_j) + \sum_{i=1}^N \alpha_i \right\} \quad (10)$$

$$\text{s.t. } \begin{cases} \sum_{i=1}^N \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C \end{cases} \quad (11)$$

for $i = 1, \dots, N$. The SVM hypothesis is given by

$$f(x) = \text{sgn} \left(\sum_{i=1}^N \alpha_i y_i K(x_i, x) + b \right). \quad (12)$$

The above formulation is not designed for an imbalanced data set where the majority class data outnumber the minority class data. A cost-sensitive version of the SVM algorithm is used in such cases, where the cost penalty parameter C in (7) is modified to weigh more to the penalties of the minority class data compared with the majority class data [25], [30]. All implementations of SVM are done using LibSVM [30]. The cost modification can be performed as follows:

$$C_i = \begin{cases} C & \text{majority class data} \\ C(r \times f) & \text{minority class data} \end{cases} \quad (13)$$

where r represents the ratio of number of majority class data to the number of minority class data and f represents a scaling factor to be tuned for a given data set.

D. Extreme Learning Machines

ELM is an emerging learning paradigm for multiclass classification and regression problems [31], [32]. An advantage of the ELM is that the training speed is extremely fast (or computationally inexpensive). The key enabler for ELMs training speed is the random assignment of input layer parameters, which do not require adaptation to the data. In such a setup, the output layer parameters can be determined analytically using an LS approach. Some of the attractive features of ELM [31] include the universal approximation capability of ELM, the convex optimization problem of ELM resulting in the

smallest training error without getting trapped in local minima, closed-form solution of ELM eliminating iterative training, and better generalization capability of ELM. ELM training involves solving the following optimization problem:

$$\min_W \left\{ \|HW - Y\|^2 + \lambda \|W\|^2 \right\} \quad (14)$$

$$H^T = \psi(W_r^T x(k) + b_r) \in \mathbb{R}^{n_h \times 1} \quad (15)$$

where λ represents the regularization coefficient, Y represents the vector of outputs or targets, ψ represents the hidden layer activation function [a sigmoidal function takes the same structure as (2)], and $W_r \in \mathbb{R}^{n \times n_h}$ and $W \in \mathbb{R}^{n_h \times 1}$ represent the input and output layer parameters, respectively. Here, n_h represents the number of hidden neurons of the ELM model, and H represents the hidden layer output matrix. The matrix W_r consists of randomly assigned elements that map the input vector to a high-dimensional feature space, while $b_r \in \mathbb{R}^{n_h}$ is a bias component assigned in a random manner similar to W_r . The number of hidden neurons determines the dimension of the transformed feature space. The elements can be assigned based on any continuous random distribution [32] and remains fixed during the learning process. Hence, the training reduces to a single-step calculation given by equation (16). The ELM decision hypothesis can be expressed as in (17). It should be noted that the hidden layer and the corresponding activation functions give a nonlinear mapping of the data, which if eliminated, becomes a linear LS model and is considered as one of the baseline models in this paper

$$W^* = (H^T H + \lambda I)^{-1} H^T Y \quad (16)$$

$$f(x) = \text{sgn} \left(W^T [\psi(W_r^T x + b_r)] \right). \quad (17)$$

The above ELM formulation is not designed to handle imbalanced or skewed data sets. As a modification to weigh the minority class data more, a simple weighting method can be incorporated in the ELM objective function (14) as

$$\min_W \left\{ (HW - Y)^T \Gamma (HW - Y) + \lambda W^T W \right\} \quad (18)$$

$$\Gamma = \begin{bmatrix} \gamma_1 & 0 & \cdot & \cdot & 0 \\ 0 & \gamma_2 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & \cdot & \cdot & \gamma_N \end{bmatrix}$$

$$\gamma_i = \begin{cases} 1 & \text{majority class data} \\ r \times f & \text{minority class data} \end{cases} \quad (19)$$

where Γ represents the weight matrix, r represents the ratio of number of majority class data to number minority class data, and f represents a scaling factor to be tuned for a given data set. This results in the training step given by (20), and the decision hypothesis takes the same form as in (17)

$$W^* = (H^T \Gamma H + \lambda I)^{-1} H^T \Gamma Y. \quad (20)$$

III. HCCI ENGINE AND DATA PROCESSING

For the purpose of identifying the stable operating envelope of HCCI engine, transient experiments are performed by exciting the engine and recording time sequences of engine variables. In this section, the HCCI engine system and experiments

TABLE I
SPECIFICATIONS OF THE EXPERIMENTAL HCCI ENGINE

Engine Type	4-stroke In-line
Fuel	Gasoline
Displacement	2.0 L
Bore/Stroke	86/86 mm
Compression Ratio	11.25:1
Injection Type	Direct Injection
Valvetrain	Variable Valve Timing with hydraulic cam phaser having 119 degree constant duration defined at 0.25mm lift, 3.5mm peak lift and 50 degree crank angle phasing authority
HCCI strategy	Exhaust recompression using negative valve overlap

performed are briefly explained followed by a methodology of labeling the data suitable for classification.

A. HCCI System and Experimentation

The concerned system of interest is a gasoline HCCI engine with a VVT system. The engine specifications are listed in Table I [13]. A schematic diagram of the experimental setup and instrumentation is shown in Fig. 1. HCCI is achieved by autoignition of the gas mixture in the cylinder. The fuel is injected early and given sufficient time to mix with air forming a homogeneous mixture. A large fraction of exhaust gas from the previous combustion cycle is retained to elevate the temperature and hence the reaction rates of the fuel and air mixture. The VVT capability of the engine enables trapping suitable quantities of exhaust gas in the cylinder.

The engine can be controlled using precalculated inputs, such as injected fuel mass (FM in mg/cyc), crank angle at intake valve opening (IVO), crank angle at exhaust valve closing (EVC), and crank angle at start of fuel injection (SOI). The valve events are measured in degrees after exhaust top dead center, while SOI is measured in degrees after combustion top dead center. Other important physical variables that influence the performance of HCCI combustion include intake manifold temperature T_{in} , intake manifold pressure P_{in} , mass flow rate of air at intake \dot{m}_{in} , exhaust gas temperature T_{ex} , exhaust manifold pressure P_{ex} , coolant temperature T_c , fuel-to-air ratio (FA), and so on. The engine performance metrics are given by combustion phasing indicated by the crank angle at 50% mass fraction burned (CA50), combustion work output given by net indicated mean effective pressure (NMEP, sometimes abbreviated as IMEP). The combustion features calculated using in-cylinder pressure, such as CA50, NMEP, and maximum rate of pressure rise R_{max} , are determined from the high-speed in-cylinder pressure measurements. The above variables at present time instant k along with their time histories are considered as inputs to the model [Section IV, (22)]. For further reading on HCCI combustion and related variables, please refer to [33].

As mentioned in Section I, the goal of this paper is to identify the HCCI operating boundary in transient operation. This requires an appropriate experiment design to obtain

transient data from the engine. A set of transient experiments is conducted at a constant speed of 2500 rotations/min and naturally aspirated conditions by varying FM, IVO, EVC, and SOI in a uniformly random manner. The experiments are conducted and data recorded using specialized engine rapid prototyping hardware. An amplitude modulated pseudorandom binary sequence (A-PRBS) has been used to design the excitation signals. A-PRBS enables exciting the engine at different amplitudes and frequencies suitable for the identification problem considered in this paper. The data are sampled using the AVL Indiset acquisition system, where in-cylinder pressure is sensed every crank angle using which the combustion features NMEP, CA50, and R_{max} are determined on a per-combustion cycle basis. More details on HCCI combustion and experiments can be found in [14]. The data are preprocessed and labeled to identify stable and unstable observations, as explained in Section III-C.

B. HCCI Instabilities

A subset of the data collected from the engine is shown in Fig. 2, where it can be observed that for some combinations of the inputs (left figures), the HCCI engine misfires (seen in the right figures where NMEP drops below 0 bar). HCCI operation is limited by several phenomena that lead to undesirable engine behavior. As described in [34], the HCCI operating range is conceptually constrained to a small region of permissible unburned (precombustion) and burned (postcombustion) charge temperature states. As previously noted, sufficiently high unburned gas temperatures are required to achieve ignition in the HCCI operating range without which complete misfire will occur. If the resulting combustion cannot achieve sufficiently high burned gas temperatures, commonly occurring in conditions with low fuel to diluent ratios or late combustion phasing, various degrees of quenching can occur resulting in reduced work output and increased hydrocarbon and carbon monoxide emissions. Under some conditions, this may lead to high cyclic variation due to the positive feedback loop existing through the trapped residual gas [10], [11]. Operation with high burned gas temperature, although stable and commonly reached at higher fueling rates where the fuel to diluent ratio is also high, yields high heat release and thus pressure rise rates that may pose challenges for engine noise and durability constraints. A discussion of the temperatures at which these phenomena occur may be found in [34].

In this paper, the considered instabilities include those modes with high cyclic variability and those with complete misfire characterized by zero work output that can be readily identified through the two aforementioned cylinder pressure-based combustion features. The other phenomena could be included with the availability of additional sensing capability or analysis methods, e.g., fast response flame ionization detection exhaust sampling equipment and detailed combustion noise analysis. Finally, it must be noted that control of these burned and unburned gas states, and therefore the potential for undesirable combustion cycles, in a recompression HCCI engine is very much a function of the engine control variables.

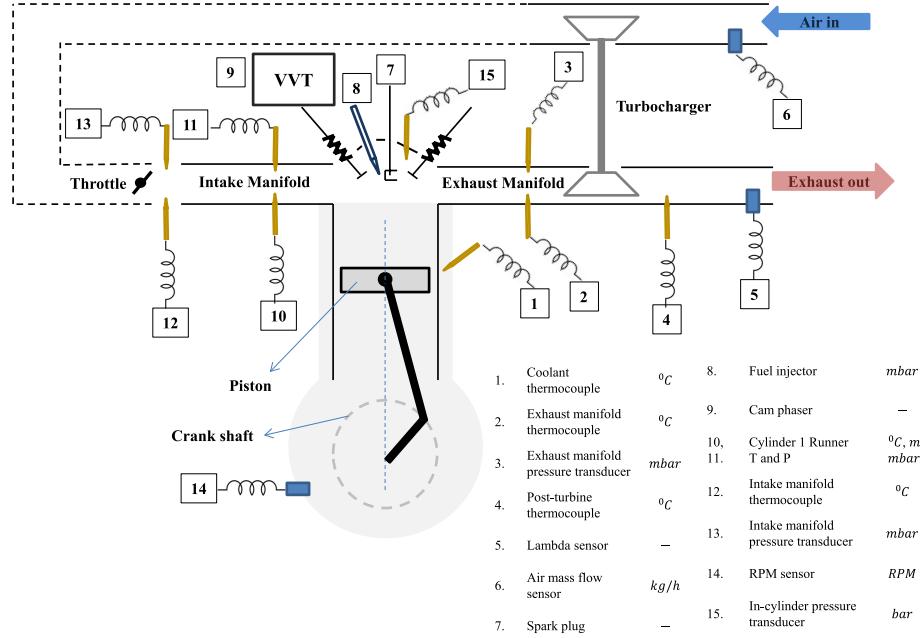


Fig. 1. Schematic diagram of the HCCI engine setup and instrumentation (only relevant instrumentation shown).

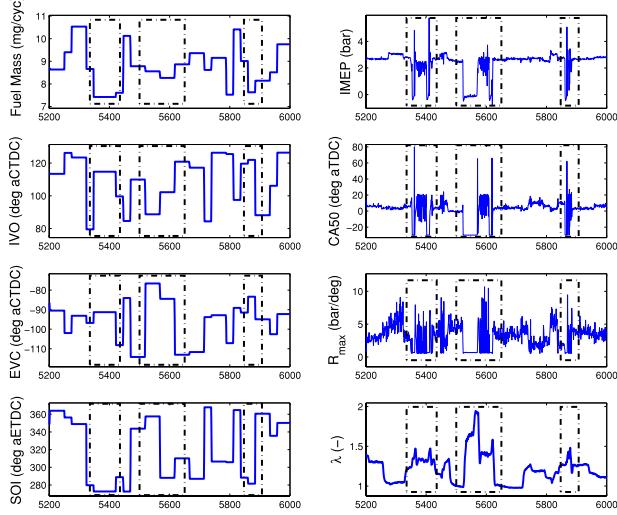


Fig. 2. Subset of the HCCI engine experimental data showing A-PRBS inputs and engine outputs. The misfire regions are shown in dotted rectangles. The data are indexed by combustion cycles.

For instance, the EVC timing will determine the trapped residual mass that will be present in the upcoming cycle, while the IVO affects both the mass of incoming air and the state of the charge during the compression stroke leading up to the autoignition. The combination of IVO and EVC (Fig. 3) defines a negative valve overlap (NVO) period where exhaust gas from the previous cycle is trapped and compressed. A larger NVO period would necessarily yield a higher trapped residual mass that would tend to increase the charge temperature and advance CA50. Likewise, the timing and mass of the fuel injection event can significantly impact the charge temperature by changing the thermodynamic properties and air-fuel ratio of the charge present during NVO.

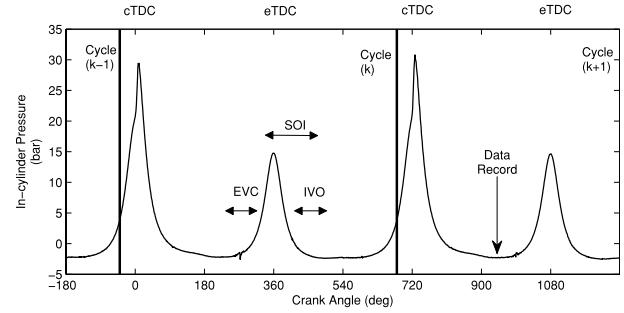


Fig. 3. Pressure trace of an HCCI combustion cycle showing valve events (IVO and EVC), injected FM, and SOI. The NVO can be seen as a smaller peak at eTDC.

The relatively high temperatures present during NVO can even lead to reactions of the fuel that will impact the temperature and chemical composition of the charge. Successful combustion of charge with a higher FM will tend to yield higher residual gas temperatures, thereby advancing CA50 in the following cycle. Likewise, an earlier SOI in NVO will tend to increase charge temperatures and reduce the ignition delay of the charge, thereby advancing CA50. As such, an improper combination of control inputs (IVO, EVC, FM, and SOI) in HCCI engines has the potential to shift operation from stable combustion to combustion with excessive heat release rates, high cyclic variability, or misfires in a single cycle.

C. Data Preprocessing and Labeling

The goal of the learning algorithm is to classify the input space into stable (future HCCI cycles are stable) or unstable (future HCCI cycles misfire or have high variability) engine behavior. The input space includes sensor measurements until

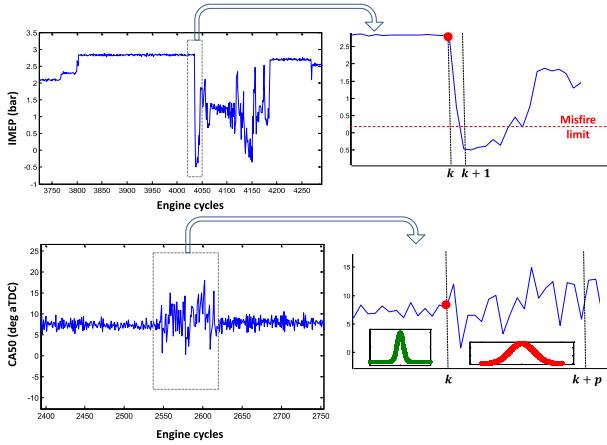


Fig. 4. Illustration showing labeling of unstable engine data.

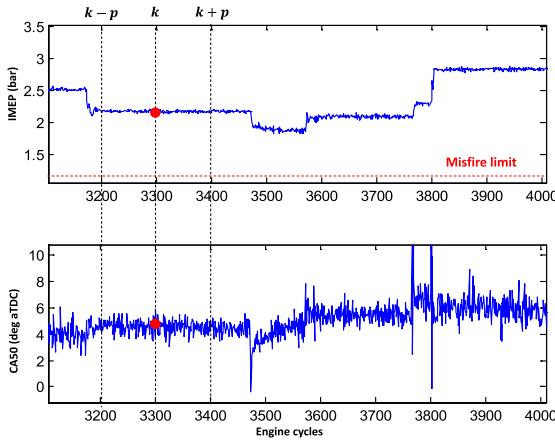


Fig. 5. Illustration showing labeling of stable engine data.

the present time instant while the indicator label depends on the future. For this purpose, the data are labeled as follows. If either of the following two conditions are met, then the data at time instant k are labeled to be unstable (Fig. 4) with a label value -1 :

- 1) an input (control inputs and past engine measurements up to an order of N_h) at cycle k results in an NMEP of less than 0.1 bar (chosen misfire limit) for any cylinder at cycle $k + 1$;
- 2) an input at cycle k results in a high variance of CA50 (any cylinder) for cycles $k + 1$ to $k + p$.

Only the first unstable data are considered in a sequence of unstable measurements. The labeling of stable data is as follows. A window of $N_w = 2p$ combustion cycles is considered (Fig. 5). If the data at cycle k are obtained as a result of stable operation in the past p cycles as well as results in stable operation in the next p cycles, it is labeled as stable observation with label value $+1$. If the time history of N_h is considered, then the data at cycle k along with the previous N_h samples are considered as inputs. If N_h is a large value, then the window length N_w can be increased accordingly. In this paper, N_w and N_h correspond to 10 and 2, respectively.

IV. MODELING THE STABLE OPERATING ENVELOPE OF HCCI ENGINE

The HCCI operating envelope is a function of the engine control inputs and engine physical variables, such as temperature, pressure, flow rate, and so on. In addition, the envelope is a dynamic system and so a predictive model requires the measurement history up to an order of N_h . The dynamic classifier model can be given by

$$\hat{y}_{k+1} = \text{sgn}(f(x_k)) \quad (21)$$

where \hat{y}_{k+1} indicates model prediction for the future cycle $k + 1$, and f can take any structure depending on the learning algorithm and x_k is given by

$$x_k = [\text{IVO}, \text{EVC}, \text{FM}, \text{SOI}, T_{\text{in}}, P_{\text{in}}, \dot{m}_{\text{in}}, T_{\text{ex}}, P_{\text{ex}}, T_c, \text{FA}, \text{NMEP}, \text{CA50}]^T \quad (22)$$

at cycle k up to cycle $k - N_h + 1$. In the following sections, the function $f(\cdot)$ is learned using the available engine experimental data.

Classification algorithms are developed based on both linear models, such as linear regression and LR, and nonlinear models, such as SVM and ELM. Several variants of the algorithms, such as data undersampling, data oversampling, no-sampling (regular data), and the cost-sensitive version are developed. The linear models (LR and LS) are compared as baselines. The engine measurements and their time histories (defined by x_k) are considered inputs to the model, while the stability labels are considered outputs. The feature vector is of dimension $n = 39$ includes sensor measurements, such as FM, IVO, EVC, SOI, T_c , T_{in} , P_{in} , \dot{m}_{in} , T_{ex} , P_{ex} , NMEP, CA50, and FA along with two cycles of history. The engine experimental data are split into training and testing sets. The training set consists of about 6400 cycles of data, while the testing set consists of about 10 200 cycles. The ratio of number of majority class data to number minority class data (r) for the training set is 17.5 and for the testing set is 16.7. The minority class data are duplicated in random so that the imbalance ratio r is one to generate the oversampled data set. The undersampled data set is generated in a similar manner by randomly removing majority class data so that $r = 1$.

A. Model Evaluation Metric

For the class imbalance problem considered here, a conventional classifier metric like the overall misclassification rate cannot be used as it would find a biased classifier, i.e., it would find a classifier that ignores the minor class data. For instance, a data set that has 95% of majority class data (with label $+1$) would achieve 95% classification accuracy by predicting all the labels to be $+1$. Hence, the following evaluation metric used for skewed data sets is considered in this paper. Let TP and TN represent the total number of positive and negative class data classified correctly by the classifier. If N^+ and N^- represent the total number of positive and negative class data, respectively, the true positive rate (TPR) and true negative rate (TNR), and the total accuracy of the classifier can be defined as follows [35]. It should be noted that the total

TABLE II
GRID SEARCH RESULTS FOR HYPERPARAMETER TUNING FOR REGULAR ELM, ELM WITH UNDERSAMPLING, AND ELM WITH OVERSAMPLING MODELS (THE MODELS RESULTING IN HIGHEST TOTAL ACCURACY ARE HIGHLIGHTED IN BOLD)

		Regular ELM					ELM with under-sampling					ELM with over-sampling				
n_h	λ	TPR					TPR					TPR				
		0.01	0.1	1	10	100	0.01	0.1	1	10	100	0.01	0.1	1	10	100
10	0.995	0.995	0.995	0.996	1.000	0.909	0.909	0.912	0.906	0.878	0.925	0.925	0.925	0.925	0.919	
	0.994	0.994	0.995	0.995	0.997	0.917	0.916	0.918	0.923	0.896	0.934	0.934	0.934	0.936	0.936	
	0.995	0.995	0.995	0.995	0.998	0.917	0.918	0.925	0.948	0.915	0.930	0.931	0.932	0.935	0.951	
	0.996	0.996	0.996	0.996	0.996	0.936	0.937	0.931	0.936	0.948	0.944	0.945	0.946	0.946	0.946	
	0.995	0.995	0.995	0.995	0.996	0.915	0.918	0.924	0.929	0.927	0.938	0.938	0.939	0.944	0.942	
30	0.333	0.333	0.327	0.258	0.000	0.732	0.734	0.737	0.771	0.743	0.716	0.716	0.716	0.722	0.757	
	0.387	0.389	0.387	0.366	0.160	0.714	0.719	0.730	0.732	0.732	0.727	0.725	0.729	0.735	0.727	
	0.430	0.430	0.423	0.407	0.294	0.771	0.773	0.752	0.735	0.727	0.773	0.773	0.771	0.755	0.724	
	0.426	0.423	0.413	0.404	0.351	0.771	0.770	0.775	0.748	0.704	0.739	0.739	0.739	0.758	0.740	
	0.433	0.430	0.420	0.405	0.356	0.789	0.794	0.784	0.768	0.755	0.763	0.768	0.773	0.775	0.775	
50	0.664	0.664	0.661	0.627	0.500	0.821	0.821	0.825	0.839	0.811	0.820	0.820	0.820	0.824	0.838	
	0.691	0.692	0.691	0.680	0.579	0.815	0.818	0.824	0.828	0.814	0.831	0.830	0.832	0.835	0.832	
	0.712	0.712	0.709	0.701	0.646	0.844	0.846	0.838	0.842	0.821	0.852	0.852	0.851	0.845	0.838	
	0.711	0.709	0.705	0.700	0.674	0.853	0.853	0.853	0.842	0.826	0.842	0.842	0.842	0.852	0.843	
	0.714	0.712	0.707	0.700	0.676	0.852	0.856	0.854	0.848	0.841	0.850	0.853	0.856	0.859	0.858	

TABLE III
GRID SEARCH RESULTS FOR HYPERPARAMETER TUNING FOR REGULAR SVM, SVM WITH UNDERSAMPLING, AND SVM WITH OVERSAMPLING MODELS (THE MODELS RESULTING IN HIGHEST TOTAL ACCURACY ARE HIGHLIGHTED IN BOLD)

		Regular SVM					SVM with under-sampling					SVM with over-sampling					
C	σ	TPR					TPR					TPR					
		0.01	0.1	1	10	100	0.01	0.1	1	10	100	0.01	0.1	1	10	100	
0.1	1.000	0.996	1.000	1.000	1.000	0.928	0.933	0.767	0.120	0.106	0.966	0.923	0.917	0.899	0.993		
	1	0.996	0.996	0.994	0.998	1.000	0.965	0.906	0.910	0.782	0.588	0.932	0.931	0.962	0.989	0.999	
	10	0.996	0.995	0.990	0.997	0.999	0.916	0.915	0.909	0.792	0.618	0.933	0.951	0.976	0.996	0.999	
	100	0.996	0.990	0.987	0.996	0.999	0.924	0.927	0.896	0.793	0.618	0.935	0.966	0.983	0.996	0.999	
	500	0.995	0.988	0.985	0.996	0.999	0.925	0.917	0.892	0.793	0.618	0.945	0.971	0.983	0.996	0.999	
1	0.000	0.374	0.000	0.000	0.000	0.632	0.735	0.931	0.998	0.998	0.667	0.820	0.913	0.958	0.162		
	1	0.397	0.423	0.552	0.108	0.054	0.668	0.825	0.923	0.967	0.987	0.792	0.884	0.814	0.221	0.082	
	10	0.423	0.444	0.645	0.145	0.080	0.802	0.882	0.925	0.958	0.975	0.822	0.817	0.732	0.167	0.080	
	100	0.430	0.567	0.650	0.165	0.080	0.833	0.886	0.915	0.954	0.975	0.848	0.763	0.642	0.165	0.080	
	500	0.436	0.627	0.637	0.165	0.080	0.853	0.882	0.910	0.954	0.975	0.833	0.724	0.623	0.165	0.080	
C	σ	Total Accuracy					Total Accuracy					Total Accuracy					
		0.01	0.1	1	10	100	0.01	0.1	1	10	100	0.01	0.1	1	10	100	
		0.500	0.685	0.500	0.500	0.500	0.780	0.834	0.849	0.559	0.552	0.816	0.871	0.915	0.928	0.577	
		1	0.696	0.710	0.773	0.553	0.527	0.817	0.866	0.916	0.875	0.787	0.862	0.907	0.888	0.605	0.540
		10	0.709	0.720	0.818	0.571	0.539	0.859	0.899	0.917	0.875	0.797	0.878	0.884	0.854	0.581	0.539
100	σ	0.713	0.779	0.819	0.580	0.539	0.879	0.906	0.905	0.874	0.797	0.892	0.865	0.812	0.580	0.539	
		500	0.716	0.808	0.811	0.580	0.539	0.889	0.899	0.901	0.874	0.797	0.889	0.847	0.803	0.580	0.539

accuracy weights the accuracy of majority and minority classes equally, i.e., only when both classes of data are classified correctly, the total accuracy value is high. This error metric

appears to be suitable for the work considered in this paper. It is desired that the envelope model is capable of predicting both stable and unstable class data with an accuracy of 90%

TABLE IV
GRID SEARCH RESULTS FOR HYPERPARAMETER TUNING OF COST-SENSITIVE SVM AND ELM MODELS
(THE MODELS RESULTING IN HIGHEST TOTAL ACCURACY ARE HIGHLIGHTED IN BOLD)

Cost-sensitive SVM							Cost-sensitive ELM							
$C \backslash \sigma$	TPR						$n_h \backslash \lambda$	TPR						
	0.001	0.01	0.1	1	10	100		10	0.909	0.909	0.909	0.907	0.901	
0.1	0.000	0.000	0.000	0.000	0.000	0.000	30	0.924	0.924	0.923	0.927	0.927		
	0.984	0.972	0.921	0.918	0.884	0.764		0.925	0.925	0.926	0.930	0.936		
	0.995	0.994	0.993	0.988	0.996	0.999		0.939	0.939	0.939	0.934	0.931		
	1.000	0.999	0.993	0.987	0.996	0.999		0.932	0.932	0.933	0.936	0.930		
	1.000	0.997	0.989	0.985	0.996	0.999		0.932	0.932	0.933	0.936	0.930		
TNR							TNR							
0.1	0.001	0.01	0.1	1	10	100	50	0.01	0.1	1	10	100		
	1.000	1.000	1.000	1.000	1.000	1.000		0.742	0.743	0.745	0.753	0.779		
	0.490	0.627	0.806	0.912	0.961	0.972		0.737	0.737	0.735	0.740	0.735		
	0.399	0.443	0.489	0.660	0.154	0.080		0.778	0.778	0.775	0.771	0.748		
	0.181	0.355	0.495	0.647	0.165	0.080		0.770	0.770	0.765	0.779	0.768		
500	0.126	0.405	0.554	0.639	0.165	0.080	90	0.786	0.786	0.794	0.784	0.784		
	Total Accuracy							Total Accuracy						
	0.001	0.01	0.1	1	10	100		0.01	0.1	1	10	100		
	0.500	0.500	0.500	0.500	0.500	0.500		0.825	0.826	0.827	0.830	0.840		
	0.737	0.800	0.863	0.915	0.922	0.868		0.830	0.830	0.829	0.833	0.831		
10	0.697	0.718	0.741	0.824	0.575	0.539	70	0.852	0.851	0.850	0.851	0.842		
	0.591	0.677	0.744	0.817	0.580	0.539		0.854	0.854	0.852	0.857	0.850		
	0.563	0.701	0.771	0.812	0.580	0.539		0.859	0.859	0.864	0.860	0.857		

and is considered the goal metric for model evaluation

$$\begin{aligned} \text{TPR} &= \frac{\text{TP}}{N^+} \\ \text{TNR} &= \frac{\text{TN}}{N^-} \\ \text{Total accuracy} &= \frac{\text{TPR} + \text{TNR}}{2}. \end{aligned} \quad (23)$$

B. Model Structure Learning

The models used for this paper (both SVM and ELM) have a very general structure that has the flexibility to map any general nonlinear function. However, it must be restricted to make the HCCI operating envelope identification task manageable. The structure of the considered nonlinear models is defined by a set of hyperparameters (cost penalty C and kernel parameter σ for SVM, while regularization coefficient λ and number of hidden neurons n_h for ELM). The hyperparameters are determined using a small proportion of the data set. In this paper, a full grid search-based cross-validation approach is employed [13]. A small subset of the training data is used to train the models using a given hyperparameter combination and its performance tested on another small subset of the training data unseen by the model. Several of such models are evaluated for minimum total accuracy and the optimal combination of hyperparameters is determined. The hyperparameter tuning results for the no-sampling and resampling cases are shown in Table II for the ELM models, while in Table III for SVM models. All the performance values reported are evaluation on unseen data sets. It can be observed that when

the data are imbalanced (no-sampling case), neither ELM nor SVM models are able to obtain good TNR accuracy. This is expected as the data are skewed and the models are biased. However, when the data are resampled, the models are able to determine the right decision boundary with a total accuracy of about 90%. Furthermore, it can be observed that the total accuracy is generally high for SVM models compared with ELM models. In addition, both sampling methods give similar accuracy levels for both ELM and SVM models. However, an advantage of undersampling can be realized in reduced computation as training is performed with a smaller subset of the training data.

The model tuning for cost-sensitive SVM and ELM is summarized in Table IV. It can be observed that the cost-sensitive models predict both positive and negative class data slightly better but without resampling the data. In addition, the total accuracy levels are slightly higher compared with undersampling and oversampling models (compare with Tables II and III). It can also be observed that the cost-sensitive algorithm gives a freedom to select models that are more accurate for one class compared with the other. By varying the scaling factor f , the boundary can be perturbed to suit the application, which requires either high TPR or high TNR. For instance, if a well-balanced model is desired, the cost-sensitive SVM with $C = 1$ and $\sigma = 1$ can be chosen. In this case, although the total accuracy is 0.915, both the positive and negative class accuracies are close to 0.91. Sensitivity plots have been shown in Figs. 6 and 7 for SVM and ELM models, respectively, to observe the variation of total accuracy with the scaling factor. By understanding the sensitivity of the weight factors,

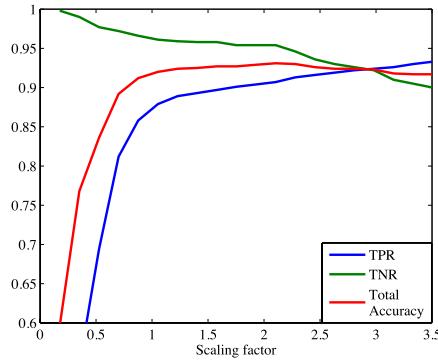


Fig. 6. Plot showing sensitivity of TPR, TNR, and total accuracy with scaling factor f for cost-sensitive SVM.

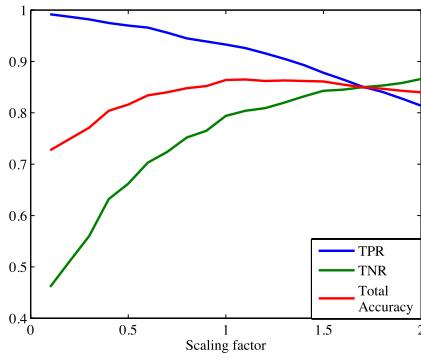


Fig. 7. Plot showing sensitivity of TPR, TNR, and total accuracy with scaling factor f for cost-sensitive ELM.

an optimal weight for the minority class data (combination of r and f) can be determined.

As mentioned earlier, the SVM models have a better total accuracy compared with the ELM models. One reason could be that SVM minimizes a hinge loss function and hence a better regularization performance. ELM, on the other hand, minimizes a squared loss function. Another reason could be that the ELM models are too simple to identify the decision boundary accurately. This can be addressed by including more hidden neurons to increase the degree of freedom of the ELM models. In addition, the ELM solution greatly depends on the random initialization of the input layer parameters (W_r and b_r). In an attempt to evaluate models with a greater number of hidden neurons and different random initializations of the input layer parameters, the following experiment is conducted. The number of hidden neurons is varied in steps of 100 up to a maximum of 1000 hidden neurons, and for each case, 10 different parameter initializations for W_r and b_r are performed using different random number generators. The results are summarized in Table V, and the model that achieved the goal accuracy of 90% with minimum number of hidden neurons is chosen as the best model (highlighted in bold). The different randomization of input layer parameters (W_r and b_r) helps in finding a compact model at a given level of accuracy (compare case 2 with case 9, where 400 additional neurons are required for a negligible improvement in total accuracy). The above analysis is performed by varying the randomization

TABLE V
COMPARISON OF COST-SENSITIVE ELM MODELS WITH DIFFERENT
RANDOM INITIALIZATION OF INPUT LAYER PARAMETERS

#	TPR	TNR	Total Accuracy	λ	n_h	f
1	0.932	0.871	0.901	10.000	800.000	1.2
2	0.921	0.879	0.900	10.000	600.000	1.3
3	0.933	0.871	0.902	10.000	900.000	1.3
4	0.930	0.863	0.896	10.000	600.000	1.2
5	0.934	0.859	0.896	10.000	800.000	1.3
6	0.927	0.871	0.899	10.000	800.000	1.2
7	0.939	0.864	0.902	10.000	800.000	1.2
8	0.927	0.869	0.898	10.000	700.000	1.1
9	0.929	0.873	0.901	10.000	1000.000	1.3
10	0.928	0.866	0.897	10.000	900.000	1.3

based on seeding the random number generator differently, and a more efficient way of initialization of input layer parameters for the HCCI engine data will be considered in the future.

C. Prediction Results

The optimal model structure (optimal hyperparameters) was determined using cross validation in the previous section. The optimal hyperparameters are fixed and the models are retrained using the complete training data set. Training involves identifying the support vectors for SVM and identifying the output layer parameters for the ELM models. The trained models are evaluated against the unseen testing data set and error metrics calculated. The models developed using SVM and ELM are compared against baseline linear models and the results summarized in Table VI. Case 2 model from Table V is considered as the best ELM model and included in Table VI.

From the modeling results summarized in Table VI, it can be observed that the SVM models (undersampling, oversampling, and cost sensitive) achieve the desired goal accuracy of 90%. This shows that both resampling methods (undersampling and oversampling) as well as cost-sensitive classification are suitable for the HCCI operating envelope problem. The nonlinear models result in better accuracies compared with the linear models indicating that the HCCI boundary is a nonlinear system and that nonlinear classification methods are indeed necessary. However, the cost paid for selecting a nonlinear model is the additional computation and memory required. As mentioned earlier, for the considered HCCI problem, it was desired that the model is capable of classifying the data with at least 90% accuracy. From Table VI, it is obvious that the decision boundary identified by linear models (underparameterized) is very simple and is not capable of achieving the target accuracy. The SVM and ELM models on the other hand require a large number of parameters in an attempt to capture complex behavior. Furthermore, the best SVM model (cost-sensitive SVM) requires about 80% more parameters for an accuracy improvement of 3% compared to the best ELM model. It should be noted that a future requirement is that the model be operated in real-time onboard the engine ECU. The engine ECU is limited in memory and computation, and hence it is concluded that the ELM model is better suited for the engine application considered in this paper.

TABLE VI
SUMMARY OF RESULTS FOR SVM AND ELM MODELS COMPARED AGAINST BASELINE LINEAR MODELS (LR AND LINEAR LS).
THE VALUE OF HYPERPARAMETERS AND NUMBER OF MODEL PARAMETERS n_p ARE ALSO INCLUDED FOR EVERY MODEL.
THE BEST SVM AND ELM MODELS ARE HIGHLIGHTED IN BOLD

	SVM				ELM				
	Regular	Under-sampling	Over-sampling	Cost-sensitive	Regular	Under-sampling	Over-sampling	Cost-sensitive	Best ELM model
TPR	0.987	0.909	0.899	0.907	0.995	0.918	0.944	0.933	0.921
TNR	0.650	0.925	0.958	0.954	0.433	0.794	0.775	0.794	0.879
Total Accuracy	0.819	0.917	0.928	0.931	0.714	0.856	0.859	0.864	0.900
λ	-	-	-	-	0.010	0.100	10.000	1.000	10.000
n_h	-	-	-	-	90.000	90.000	90.000	90.000	600.000
f	-	-	-	2.104	-	-	-	0.909	0.769
C	100	10	0.1	1	-	-	-	-	-
σ	1	1	10	10	-	-	-	-	-
n_p	33696	16965	236691	120900	3690	3690	3690	3690	24600

	Logistic Regression				Linear LS				
	Regular	Under-sampling	Over-sampling		Regular	Under-sampling	Over-sampling	Cost-sensitive	
TPR	0.995	0.911	0.928		0.996	0.941	0.955	0.875	
TNR	0.441	0.791	0.786		0.389	0.704	0.699	0.828	
Total Accuracy	0.718	0.851	0.857		0.692	0.822	0.827	0.852	
n_p	40	40	40		40	40	40	40	

D. Model Interpretation

The best ELM and SVM models identified in the previous section are used to make predictions on unseen engine inputs and predictions are shown in Figs. 8 and 9, respectively, while quantitative results are included in Table VI. As mentioned earlier, the operating envelope is a decision boundary in the input space within which any input operates the HCCI in a stable manner and any input outside the envelope might operate the engine in an unstable manner. The models are evaluated as follows. Using the available engine sensors, the HCCI variables, such as NMEP, CA50, and FA, along with engine control inputs, such as FM, IVO, EVC, and SOI, and engine physical states, such as T_{in} , P_{in} , \dot{m}_{in} , T_{ex} , P_{ex} , and T_c , at time instant k are given as input to the models (22). The model predictions at time $k + 1$ are obtained using (17) for ELM model and (12) for the SVM model. The engine's actual response at time $k + 1$ is also recorded. A data point is marked in red if the model predicts the engine operation to be unstable (-1), while it is marked in green if the model predicts

the data point to be stable ($+1$). In the figures, a dotted line in the NMEP plot indicates the misfire limit, a dotted ellipse in CA50 plot indicates high-variability instability mode, while a dotted rectangle indicates misclassified predictions by model. The input feature dimension is about 39 and it is impossible to show the effect of all inputs on the model prediction. However, the predictions are imposed on NMEP and CA50 plots as the stability of HCCI operation was evaluated using these two quantities in this paper. For an additional perspective on changes in control inputs, the fueling input (defined as FM earlier) is also included in the plots. It should be understood that FM is not the only input for prediction and the signals are defined as in (22).

It can be seen from the above plots that as a whole, both SVM and ELM models classify the HCCI engine data fairly well in spite of the high-amplitude noise inherent in the HCCI experimental data. The data consist of step changes in FM, IVO, EVC, and SOI, and whenever a bad combination of inputs is chosen, the engine either misfires completely (see NMEP fall below misfire limit) or exhibits high-variability

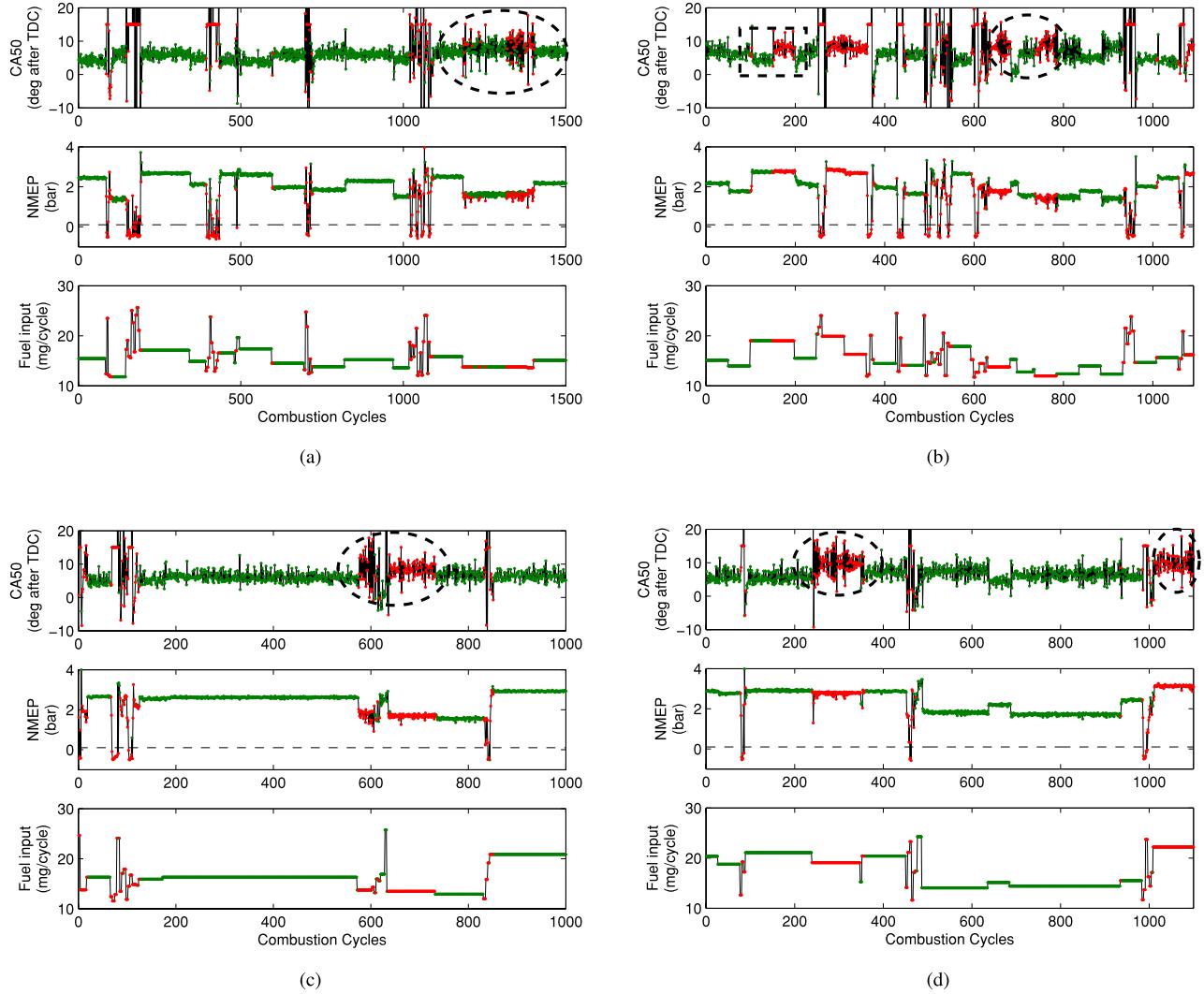
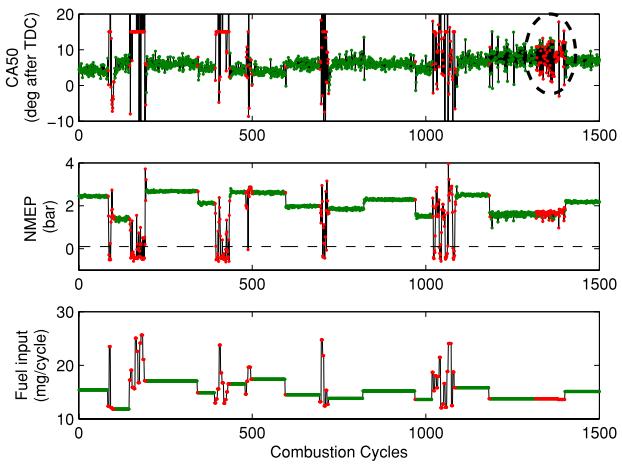


Fig. 8. Prediction results of the best ELM envelope model imposed on NMEP, CA50, and one input variable (FM) for four unseen data sets (subplots a, b, c and d). The color code indicates model prediction—green (and red) indicates the model predicting the engine state to be stable (and unstable). The dotted lines in the NMEP plot indicate misfire limit, the dotted ellipses in CA50 plot indicate high-variability instability mode, while the dotted rectangles indicate misclassified data observations.

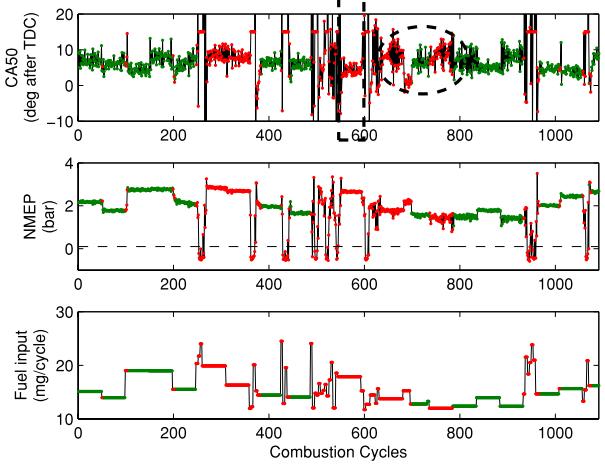
combustion (see dotted ellipses). The goal of this paper as stated previously is to predict if a future HCCI combustion event is stable or unstable based on available measurements. The results summarized in Table VI are proof that the developed models indeed accomplished the goal with a target accuracy of 90%. The remaining 10% of the data set is misclassified [see dotted rectangle in Fig. 8(b) for ELM model and in Fig. 9(b) for SVM model to observe a few misclassified observations]. The reason for misclassification could be either the model overfitting the training data or a lack of sufficient data in certain regions of the input space leading to a poor approximation.

In order to get a closer look at the predictions, zoomed-in plots of the ELM envelope model's prediction are shown in Fig. 10. It can be observed that when a step change in control inputs is made at time k to $k + 1$, the model receives information about the state of the engine at time instant k , but the actuator change at $k + 1$ if known can be used to make predictions about the future cycle $k + 1$. For instance, Fig. 10(a)

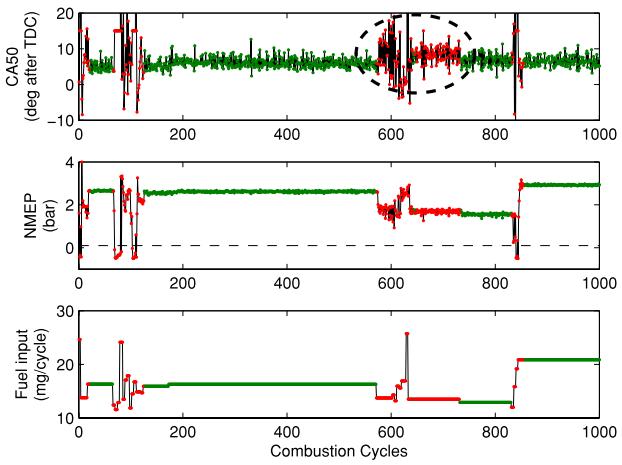
shows that up to cycle 680, the engine exhibits unstable combustion, which is predicted correctly by the model. In addition, the actuator step changes at cycle 681 is used along with engine measurements at cycle 680 to predict the nature of HCCI at cycle 681. It can be observed that the model correctly predicts the stable nature of HCCI for this particular set point. In addition, Fig. 10(b) shows a misfire situation around cycles 950 where several set points on fueling (and other actuators not shown in the figure) are applied on the engine, but since these inputs fall outside the operating envelope, the engine misfires and the model predicts it correctly as well. This clearly shows that the developed model is a representation of the dynamic operating envelope of the engine, i.e., it predicts if the engine operation is stable/unstable at time $k + 1$, given the measurements up to time k . Hence, using the model, any given input actuator setting and history of measurements of key combustion variables, it would be possible to predict if the subsequent set of combustion cycles misfire or not. It should be noted that the model is about 90% accurate and wrong predictions



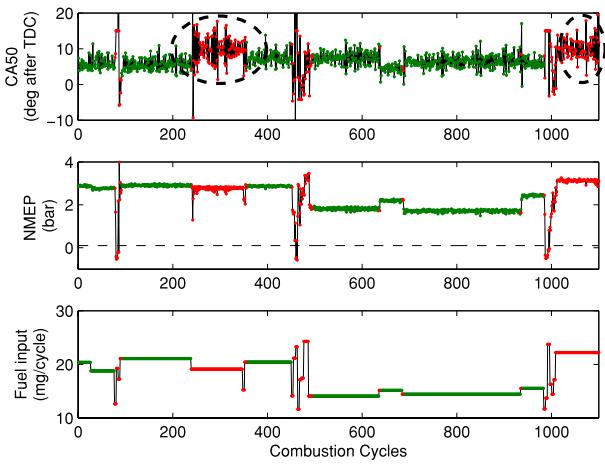
(a)



(b)

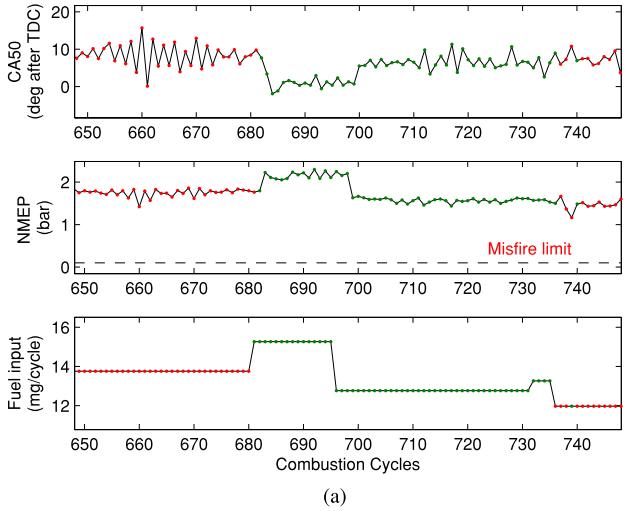


(c)

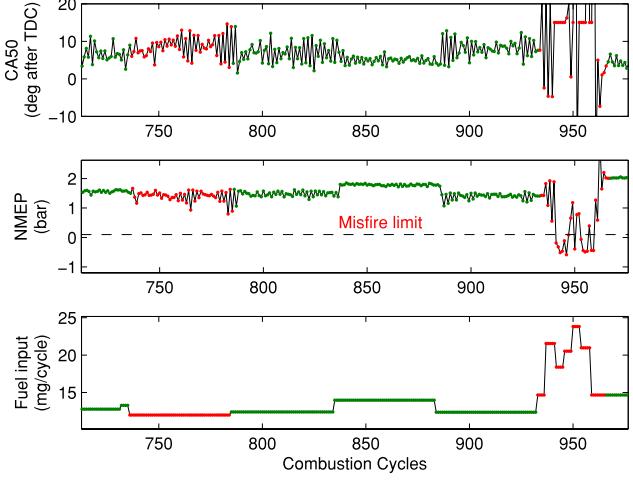


(d)

Fig. 9. Prediction results of the cost-sensitive SVM envelope model imposed on NMEP, CA50, and one input variable (FM) for four unseen data sets (subplots a, b, c and d). The color code indicates model prediction—green (and red) indicates the model predicting the engine state to be stable (and unstable). The dotted lines in the NMEP plot indicate misfire limit, the dotted ellipses in CA50 plot indicate high-variability instability mode, while the dotted rectangles indicate misclassified data observations.



(a)



(b)

Fig. 10. Zoomed-in plots for two unseen data sets (subplots a and b) showing prediction results of the best ELM envelope model imposed on NMEP, CA50, and one input variable (FM). The color code indicates model prediction—green (and red) indicates the model predicting the engine state to be stable (and unstable).

are to be expected from both SVM and ELM models owing to insufficient data or possible overfitting. Further analysis is to be conducted to study the confidence of such models in

different operating regions of the engine. The model developed in this paper is ideal for predictive analysis, fault detection, and control onboard the engine ECU. Furthermore, the presented

approach is novel and generic that can be applied for different engines and different definitions of stability.

V. CONCLUSION

HCCI engines have a narrow region of stable operation and it is important to gain knowledge about the operating envelope for diagnostics and controls development. In this paper, a novel solution using classification learning has been developed that predicts the stability of future combustion events based on past and present measurements of the engine along with control inputs. An imbalanced classification problem has been formulated and solved using linear and nonlinear methods, such as LR, linear regression, SVM, and ELM. A comparison of data resampling methods and cost-sensitive learning approaches has been performed and the results are summarized. Resampling methods are found to work well but cost-sensitive methods have a slightly better accuracy, and more importantly, avoid artificial modifications in the data distributions. A modification to the ELM algorithm has been made by weighting the minority class data more to handle the imbalance in the data set. The cost-sensitive SVM classifier outperforms the linear and ELM models in terms of accuracy but requires a large fraction of the data to be stored for predictions. The SVM envelope model requires about 80% more parameters for an accuracy improvement of 3% compared with the ELM envelope model. Although the ELM envelope model is slightly inferior in terms of accuracy, it is preferred over SVM as it is much simpler (less number of parameters) and appears to be more suitable for online implementation on a memory-limited system, such as the HCCI engine ECU. Future work would focus on identifying more unstable modes of HCCI and performing online model adaptation to improve prediction accuracy. Furthermore, applications of the developed model toward designing efficient diagnostics and control systems would be considered in the future.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

REFERENCES

- [1] R. Thring, *Homogeneous-Charge Compression-Ignition Engines*. Warrendale, PA, USA: SAE, 1989.
- [2] M. Christensen, P. Einewall, and B. Johansson, "Homogeneous charge compression ignition using iso-octane, ethanol and natural gas-a comparison to spark ignition operation," in *Proc. Int. Fuels, Lubricants Meeting, Exposit.*, Tulsa, OK, USA, Oct. 1997, pp. 1–3, paper 972874.
- [3] T. Aoyama, Y. Hattori, J. Mizuta, and Y. Sato, "An experimental study on premixed-charge compression ignition gasoline engine," in *Proc. Int. Congr., Exposit.*, Detroit, MI, USA, Feb. 1996, pp. 1–10, paper 960081.
- [4] K. Chang, A. Babajimopoulos, G. A. Lavoie, Z. S. Filipi, and D. N. Assanis, *Analysis of Load and Speed Transitions in an HCCI Engine Using 1-D Cycle Simulation and Thermal Networks*. Warrendale, PA, USA: SAE, 2006.
- [5] J. Bengtsson, P. Strandh, R. Johansson, P. Tunestal, and B. Johansson, "Model predictive control of homogeneous charge compression ignition (HCCI) engine dynamics," in *Proc. IEEE Int. Conf. Control Appl.*, Oct. 2006, pp. 1675–1680.
- [6] Y. Wang, S. Makkapati, M. Jankovic, M. Zubeck, and D. Lee, *Control Oriented Model and Dynamometer Testing for a Single-Cylinder, Heated-Air HCCI Engine*. Warrendale, PA, USA: SAE, 2009.
- [7] C.-J. Chiang, A. Stefanopoulou, and M. Jankovic, "Nonlinear observer-based control of load transitions in homogeneous charge compression ignition engines," *IEEE Trans. Control Syst. Technol.*, vol. 15, no. 3, pp. 438–448, May 2007.
- [8] Y. Urata, M. Awasa, J. Takanashi, T. Kakinuma, T. Hakozaki, and A. Umemoto, *A Study of Gasoline-Fuelled HCCI Engine Equipped with an Electromagnetic Valve Train*. Warrendale, PA, USA: SAE, Jun. 2004.
- [9] R. Scarunge, C. Wildman, and W. K. Cheng, "On the high load limit of boosted gasoline HCCI engine operating in NVO mode," *SAE Int. J. Engines*, vol. 3, pp. 35–45, Apr. 2010.
- [10] G. T. Kalghatgi and R. A. Head, "Combustion limits and efficiency in a homogeneous charge compression ignition engine," *Int. J. Engine Res.*, vol. 7, no. 1, pp. 215–236, 2006.
- [11] M. Shahbakhti and C. R. Koch, "Characterizing the cyclic variability of ignition timing in a homogenous charge compression ignition engine fueled with n-heptane/iso-octane blend fuels," *Int. J. Engine Res.*, vol. 9, pp. 361–397, Jan. 2008.
- [12] I. V. Kolmanovsky and E. G. Gilbert, *Support Vector Machine-Based Determination of Gasoline Direct Injected Engine Admissible Operating Envelope*. Warrendale, PA, USA: SAE, Mar. 2002.
- [13] V. M. Janakiraman, X. Nguyen, and D. Assanis, "Nonlinear identification of a gasoline HCCI engine using neural networks coupled with principal component analysis," *Appl. Soft Comput.*, vol. 13, no. 5, pp. 2375–2389, 2013.
- [14] V. M. Janakiraman, X. Nguyen, J. Sterniak, and D. Assanis, "A system identification framework for modeling complex combustion dynamics using support vector machines," in *Informatics in Control, Automation and Robotics* (Lecture Notes in Electrical Engineering), vol. 283. New York, NY, USA: Springer-Verlag, 2014, pp. 297–313.
- [15] A. Soliman, G. Rizzoni, and V. Krishnaswami, *The Effect of Engine Misfire on Exhaust Emission Levels in Spark Ignition Engines*. Warrendale, PA, USA: SAE, Feb. 1995.
- [16] P. Azzoni, D. Moro, C. M. Porceddu-Cilione, and G. Rizzoni, *Misfire Detection in a High-Performance Engine by the Principal Component Analysis Approach*. Warrendale, PA, USA: SAE, 1996.
- [17] C. J. Brace, S. Akehurst, and M. Ward, "The use of prior knowledge to accelerate the determination of the permissible operating envelope of an internal combustion engine," *Inst. Mech. Eng., D, J. Autom. Eng.*, vol. 225, no. 2, pp. 206–221, 2011.
- [18] J.-D. Wu and C.-H. Liu, "An expert system for fault diagnosis in internal combustion engines using wavelet packet transform and neural network," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 4278–4286, Apr. 2009.
- [19] E. Ftoutou, M. Chouchane, and N. Besbès, "Feature selection for diesel engine fault classification," in *Condition Monitoring of Machinery in Non-Stationary Operations*. Berlin, Germany: Springer-Verlag, 2012, pp. 309–318.
- [20] D. Lee and G. Rizzoni, *Detection of Partial Misfire in IC Engines Using a Measurement of Crankshaft Angular Velocity*. Warrendale, PA, USA: SAE, 1995.
- [21] D. Lee and G. Rizzoni, *Misfire Pattern Recognition in High Performance SI 12-Cylinder Engine*. Warrendale, PA, USA: SAE, 1996.
- [22] V. M. Janakiraman, X. Nguyen, and D. Assanis, "A Lyapunov based stable online learning algorithm for nonlinear dynamical systems using extreme learning machines," in *Proc. IJCNN*, 2013, pp. 1–8.
- [23] B. Meng, Y. Wang, and Y. Yang, "Efficiency-optimization control of extended range electric vehicle using online sequential extreme learning machine," in *Proc. IEEE VPPC*, Oct. 2013, pp. 1–6.
- [24] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
- [25] K. Veropoulos, C. Campbell, and N. Cristianini, "Controlling the sensitivity of support vector machines," in *Proc. Int. Joint Conf. AI*, 1999, pp. 55–60.

- [26] A. Dobson, *An Introduction to Generalized Linear Models* (Texts in Statistical Science Series), 2nd ed. New York, NY, USA: Taylor & Francis, 2010.
- [27] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag, 1995.
- [28] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," *Adv. Neural Inf. Process. Syst.*, vol. 9, pp. 155–161, Jan. 1996.
- [29] B. Schölkopf, P. Bartlett, A. Smola, and R. Williamson, "Support vector regression with automatic accuracy control," in *Proc. ICANN*, 1998, pp. 111–116.
- [30] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.
- [31] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, pp. 489–501, May 2006.
- [32] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 42, no. 2, pp. 513–529, Apr. 2012.
- [33] F. Zhao, T. N. Asmus, D. N. Assanis, J. E. Dec, J. A. Eng, and P. M. Najt, *Homogeneous Charge Compression Ignition (HCCI) Engines*. Warrendale, PA, USA: SAE, Mar. 2003.
- [34] G. A. Lavoie, J. Martz, M. Wooldridge, and D. Assanis, "A multi-mode combustion diagram for spark assisted compression ignition," *Combustion Flame*, vol. 157, no. 6, pp. 1106–1110, 2010.
- [35] K.-A. Toh, "Deterministic neural classification," *Neural Comput.*, vol. 20, no. 6, pp. 1565–1595, Jun. 2008.



Vijay Manikandan Janakiraman received the bachelor's degree in mechanical engineering from the Sri Venkateswara College of Engineering, Chennai, India, the master's degrees in mechanical engineering and electrical engineering (systems), and the Ph.D. degree in mechanical engineering from the University of Michigan, Ann Arbor, MI, USA, in 2007, 2008, 2013, and 2013, respectively.

He has been a Research Scientist with the Data Sciences Group, Intelligent Systems Division, NASA Ames Research Center, Moffett Field, CA, USA, since 2013. His current research interests include machine learning and data mining in high-dimensional time series, dynamical systems, optimization, and control.



XuanLong Nguyen received the Ph.D. degree in computer science and the M.S. degree in statistics from the University of California at Berkeley (UC Berkeley), Berkeley, CA, USA.

He is currently an Assistant Professor of Statistics with the University of Michigan, Ann Arbor, MI, USA. His current research interests include distributed and variational inference, nonparametric Bayesian statistics, and applications to detection/estimation problems in distributed and adaptive systems.

Dr. Nguyen is a recipient of the CAREER Award from the NSF Division of Mathematical Sciences, the Leon O. Chua Award from the UC Berkeley, the IEEE Signal Processing Society Young Author Best Paper award, and an Outstanding Paper Award from the International Conference on Machine Learning.



Jeff Sterniak received the B.S. degree in computer engineering and the M.S.E. degree in automotive systems engineering from the University of Michigan, Dearborn, MI, USA.

He is currently the Combustion and Emissions Team Lead with the System and Advanced Engineering Group, Gasoline Systems Division, Robert Bosch LLC, Farmington, MI, USA. His current research interests include combustion system optimization for electrified powertrains and system concepts for reducing particulate emissions.



Dennis Assanis received the Ph.D. degree in power and propulsion and the M.S. degrees in naval architecture and marine engineering and mechanical engineering from the Massachusetts Institute of Technology, Cambridge, MA, USA.

He is a Professor with the Department of Mechanical Engineering and the Provost, Senior Vice President for Academic Affairs, and Vice President for Brookhaven Affairs with the Stony Brook University, NY, USA. He served as the Jon R. and Beverly S. Holt Professor of Engineering and Arthur F. Thurnau Professor with the University of Michigan, Ann Arbor, MI, USA, as well as the Director of the Michigan Memorial Phoenix Energy Institute, University of Michigan, Founding Director with the U.S.-China Clean Energy Research Center for Clean Vehicles, and the Director of the Walter E. Lay Automotive Laboratory, Ann Arbor. His current research interests include the thermal sciences and their applications to energy conversion, power and propulsion, and automotive systems design. His research focuses on analytical and experimental studies of the thermal, fluid, and chemical phenomena that occur in internal combustion engines, after-treatment systems, and fuel processors. His efforts to gain new understanding of the basic energy conversion processes have made significant impact in the development of energy and power systems with significantly improved fuel economy and dramatically reduced emissions. His group's research accomplishments have been published in more than 250 articles in journals and international conference proceedings.

Dr. Assanis is a member of the National Academy of Engineering and a fellow of ASME and SAE.