Prediction of the Derived Cetane Number of Hydrocarbon Fuels Using Extended-Wavelength FTIR Spectra and Support Vector Regression

Vivek Boddapati, Alison M. Ferris, Ronald K. Hanson
Department of Mechanical Engineering, Stanford University
Stanford, CA, USA

1 Introduction

The need for rapid mitigation of climate change and aviation-related emissions has been the driving force behind the development and deployment of alternative jet fuels (AJFs) [1]. Research efforts have mainly focused on hydrocarbons (and their blends) derived from non-petroleum feedstocks, designed to act as a direct replacement for conventional distillate jet fuels in existing engine/aircraft systems. A key challenge associated with the current AJF approval process is the need for rigorous tests to ensure the safe operability and performance of these fuels, often requiring large volumes of the novel fuel candidates. Cetane number is one of the fuel properties that is important for characterizing AJFs due to its practical relevance as an ignition quality measure. The cetane number of a fuel is typically determined by comparing its combustion characteristics in a test engine with those of reference fuels (ASTM D613 [2]). However, this method requires a large volume of fuel on the order of 1 L. An alternative metric is the derived cetane number (DCN), which is estimated based on a fuel's ignition delay time (IDT) measured in an ignition quality tester (ASTM D6890 [3]). Like the cetane number, DCN also serves as an effective index for ignition quality, but has the advantage of a lower fuel volume requirement (around 100 mL) and relatively smaller reproducibility errors. Even with the reduced volume requirement, this DCN measurement still poses a hindrance since it is only one among numerous other property tests that are needed for approving AJFs.

Consequently, a variety of low-volume, low-cost prescreening techniques have been identified in order to predict fuel properties and model their combustion behavior before entering the official AJF certification process, as described in [4]. These techniques aim to develop correlations between fuel properties and compositional information obtained from methods such as 2-D gas chromatography (GCxGC), nuclear magnetic resonance (NMR), or Fourier transform infrared (FTIR) spectroscopy. Various studies over the years have reported such correlations for estimating DCN. Ure and Dooley [5] developed numerical methods based on NMR for predicting the derived cetane number (DCN) of hydrocarbons, while Yang et al. [6] leveraged GCxGC data to predict the DCN and other critical properties of sustainable aviation fuels using a Monte Carlo sampling technique. A few other studies proposed ways to correlate cetane number with infrared spectral data. Cooper et al. [7] used near-infrared (NIR) spectral data (880–1570 nm) of liquid fuels to build partial least squares (PLS) models for predicting the cetane index, among

Correspondence to: vivekb98@stanford.edu

other physical properties, of jet fuels. Wang et al. [8] proposed the vapor-phase infrared absorbance ratio at two wavelengths (3.41 μm/3.39 μm) as a spectroscopic predictor for the DCN of hydrocarbon fuels. More recently, Al Ibrahim and Farooq [9] used IR spectra to develop three methods for predicting the DCN of diesel-like fuels: PLS, support vector machine (SVM), and artificial neural network (ANN).

However, the datasets used to train these models are usually restricted to neat hydrocarbons and conventional jet fuels, thus limiting their applicability to novel AJF candidates. Our recent work [10] at Stanford proposed the use of elastic-net regularized linear models based on FTIR spectral data over an extended wavelength range of 2-15.38 μm (henceforth referred to as the 2-15 μm region) for the prediction of three combustion properties (derived cetane number, ignition delay time, and net heat of combustion). The training dataset used in this previous work included the vapor-phase FTIR spectra and property data of approved AJFs along with those of conventional jet fuels and neat hydrocarbons, and were shown to achieve high prediction accuracy on test fuel blends comprised of AJFs. Although the performance of this linear model for DCN was satisfactory, it underperformed relative to the other properties that were studied, indicating that the spectrum-DCN relationship could be better captured using nonlinear regression methods. Support vector regression (SVR) is one such method that offers the capability to effectively model nonlinear property correlations, without being as computationally expensive as more complex methods such as artificial neural networks (ANN). SVR is also less prone to overfitting than ANN when the training dataset has less samples than features (as is the case with FTIR spectra). Despite having reduced interpretability compared to linear models, SVR has the potential to predict key properties such as DCN with much higher accuracy. In the present work, we develop a nonlinear support vector regression model for predicting DCN based on extended-wavelength FTIR spectra, and compare its performance with our previously developed elastic-net regularized linear model [10] on a wide range of hydrocarbon fuels, including AJFs.

2 Methodology

2.1 Training Dataset of FTIR Spectra and DCN Values

The training dataset used in the development of the current models contains the vapor-phase FTIR spectra of 149 fuels in the 2-15 μm wavelength range, spanning carbon numbers from 5 to 16. Out of these, the spectra of 34 neat hydrocarbons (n-, iso-, cyclo-paraffins, and aromatics) were sourced from the Pacific Northwest National Laboratory (PNNL) spectral database [11], while the spectra of 13 neat hydrocarbons, 8 conventional jet fuels (Jet A) and 9 approved AJFs (C fuels) were measured at Stanford University using a Nicolet 6700 FTIR spectrometer (details of the measurement procedure can be found in [12]). The spectra of 85 blends of neat hydrocarbons were calculated as the mole fraction-weighted sums of the individual components' spectra.

The dataset also contains DCN values compiled from various literature and online sources. To maintain consistency, all the DCN values used in the dataset were taken from experimental measurements adhering to ASTM D6890 [3]. The DCN data for neat hydrocarbons and jet fuels were sourced from Yanowitz et al. [13] and Wang et al. [8,14], respectively. The DCN values of the blends of neat hydrocarbons were sourced from [8, 15].

2.2 Model Development

The FTIR spectra in the 2-15 μm wavelength range contain more than 18,000 wavelengths, while the total number of fuels in the current training dataset is only 149. Such high dimensionality in the training

dataset usually leads to poor performance of machine learning models. In order to circumvent the problem of high dimensionality, the FTIR spectra are first pre-processed by performing principal component analysis (PCA). The FTIR data are linearly transformed into a set of orthogonal principal components (PCs), such that the first PC captures the greatest variance in the original dataset, followed by the second PC, and so on. This enables the selection of a subset of PCs as predictors for training regression models, thereby greatly reducing the dimensionality without significant loss of information. The number of principal components to select, N_{PC}, is treated as a hyperparameter that can be optimized using cross-validation. SVR in itself is a linear regression method. However, SVR can be employed for nonlinear regression when used in conjunction with a kernel function that maps the original input data to a higher-dimensional feature space, where the problem is linear. In the present work, the selected set of PCs is similarly mapped to a high-dimensional space using a radial basis function (RBF) kernel, parameterized by σ. Other kernel functions such as linear and polynomial were found to underperform compared to RBF on the current dataset. An SVR model is then trained on the transformed features to obtain the best-fit hyperplane for predicting the property of interest. The SVR algorithm has an additional tunable cost parameter C, which controls the penalty on outlier data points. Further details about the SVR approach can be found in [16, 17].

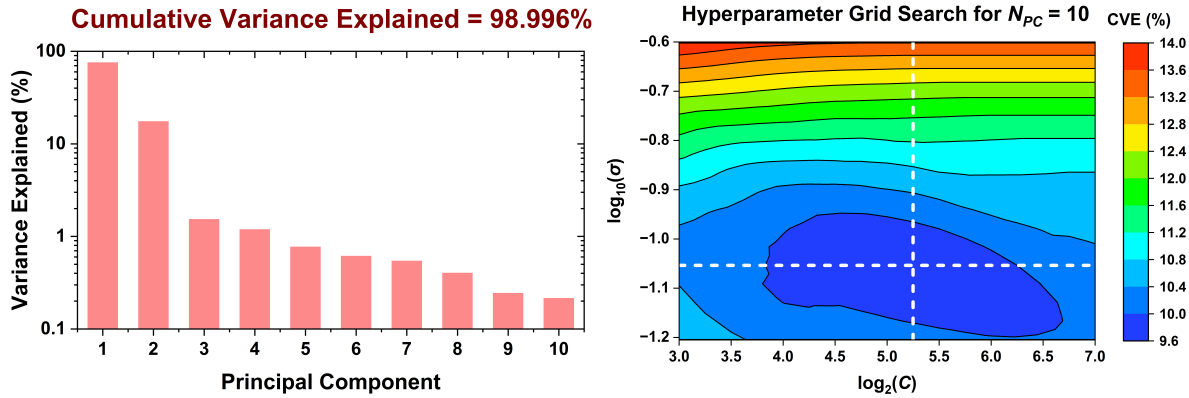

Figure 1: Left: Percentage of variance explained by each of the first 10 principal components. Right: Cross-validated grid search for selecting the optimal values of the model parameters C and σ (indicated by the dashed white lines).

To achieve optimal performance, the best combination of parameters N_{PC}, C, and σ needs to be identified. This is done by performing a 10-fold cross-validated grid search. By experimenting with different values of these three parameters, the model selects the combination that results in the minimum cross-validation error (CVE). Following this process, only the first ten PCs are chosen to build the SVR model for DCN. As seen in the left panel of Fig. 1, the first 10 PCs are able to cumulatively capture close to 99% of the total variance in the FTIR dataset. Thus, discarding the subsequent PCs causes virtually no loss of spectral information. The right panel of Fig. 1 demonstrates the model tuning process over a grid of C and σ values, with N_{PC} fixed at its optimal value of 10. The dashed white lines correspond to the combination of C and σ that has the lowest CVE. Therefore, the optimal parameters describing the SVR model for DCN are: $N_{PC} = 10$, $C = 38.05$, and $\sigma = 0.088$.

3 Results and Discussion

The optimal hyperparameters obtained using the procedure described in Section 2.2 were used to generate the SVR model for DCN. The performance of this optimized model was evaluated on the training dataset using three performance metrics – cross-validation error (CVE, an estimate of future prediction

error), coefficient of determination (R^2, a measure of goodness-of-fit), and root-mean-squared error (RMSE) of prediction. The results from the current SVR model are also compared with the previously developed elastic-net regularized linear model [10] to highlight the improvement in predictive performance due to the use of nonlinear regression. Figure 2 shows the performance of the previous (linear) and current (nonlinear) models on the training dataset for DCN. Figs. 2a and 2b show the predicted versus actual DCN and the corresponding residuals obtained using the linear model, while Figs. 2c and 2d show the predicted versus actual DCN and residuals using the nonlinear model. The CVE, R^2, and RMSE for these models are also listed in the figure headings.

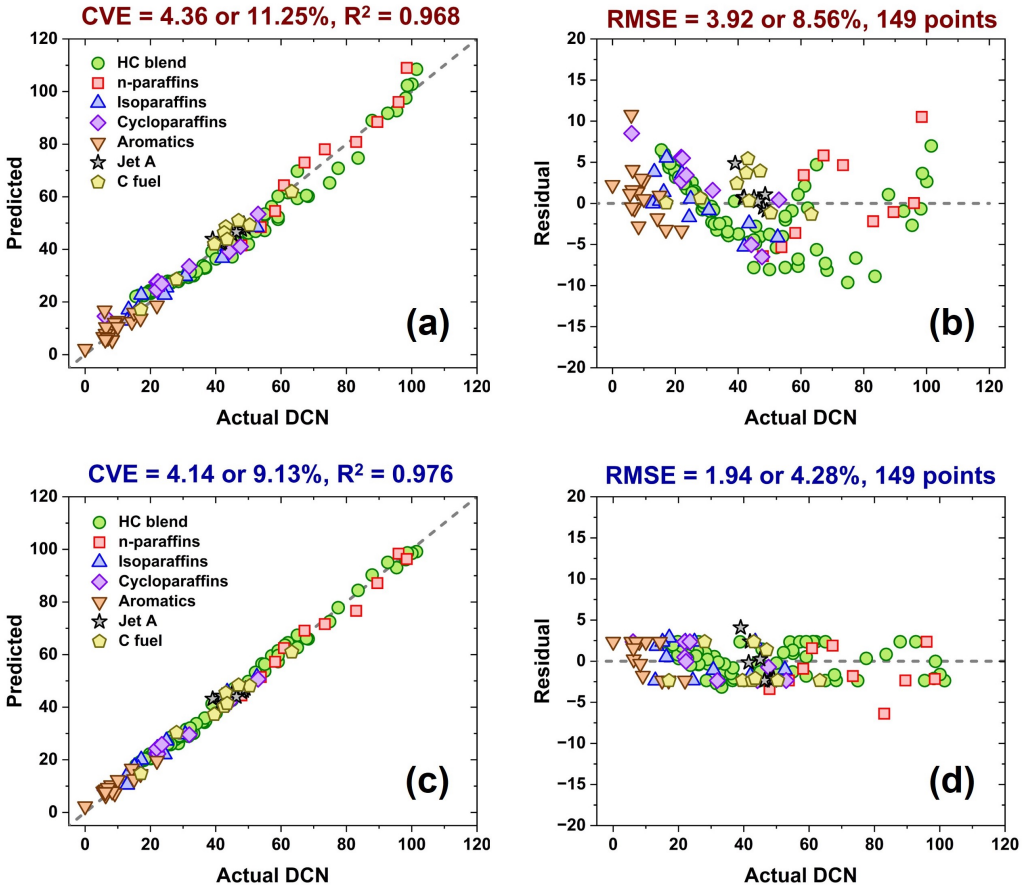

Figure 2: DCN model performance on training data: (a) Predicted DCN and (b) residuals using the previously developed elastic-net regularized linear model [10]; (c) Predicted DCN and (d) residuals using the current SVR model.

The nonlinear model for DCN shows distinctively improved predictive performance compared to the linear model, as evidenced by the reduction in CVE and RMSE, and the increase in R^2. Although the CVE of the SVR model is only 5% lower than the linear model, there is a factor of 2 reduction in the RMSE achieved. Additionally, the higher R^2 value for the current model indicates a stronger correlation between FTIR spectra and DCN. Direct comparison of these models with alternative models proposed in the literature is not straightforward, mainly because of the differences in the training data and the reported performance metrics. However, our SVR model for DCN (R^2 = 0.976, CVE = 9.13%) has lower percentage error than other PLS models based on IR spectra (R^2 = 0.91, CVE = 14.03%), and achieves superior performance compared to similar SVM-based models (R^2 = 0.93, CVE = 10.73%) [9]. Since the present model is aimed at streamlining the AJF approval process, comparing its prediction accuracy with standard test methods is a crucial step in evaluating model performance. The reproducibility error

of the ASTM D6890 test method is estimated to be equal to $0.06201(\text{DCN} - 13.7)$ [3]. The root-mean-squared reproducibility error calculated using this formula on the current training dataset is found to be 2.13 or 4.70%, which is higher than the RMSE achieved by the SVR model for DCN (1.94 or 4.28%). This indicates that the uncertainty associated with the experimental DCN values in the training dataset is expected to impact the uncertainty in the predicted DCN values to a greater extent than the modeling approach itself.

4 Conclusions

The vapor-phase FTIR spectra of jet fuel-relevant hydrocarbon fuels in the 2-15.38 μm wavelength range were used to develop a nonlinear regression model for predicting derived cetane number. The FTIR spectra were first pre-processed using principal component analysis (PCA), and a support vector regression (SVR) model was trained on a subset of the principal components to reduce the dimensionality of the dataset. The optimal model parameters were chosen by performing a cross-validated grid search. The results from this optimal SVR model were compared with the elatic-net regularized linear model developed in our previous work at Stanford [10]. The present nonlinear model showed considerable improvement in performance compared to the previous linear model, and in general, performed better than other predictive methods proposed in the literature. The prediction error achieved by the SVR model was also found to be lower than the reproducibility error of the standard test method for estimating DCN (ASTM D6890 [3]), highlighting the dominance of experimental uncertainty in the training data over the prediction uncertainty associated with the modeling approach. Future work aims to implement this support vector regression approach for modeling other important physical and chemical properties of hydrocarbon-based, alternative fuels.

5 Acknowledgements

This research was funded by the U.S. Federal Aviation Administration Office of Environment and Energy through ASCENT, the FAA Center of Excellence for Alternative Jet Fuels and the Environment, Project 25 through FAA Award Number 13-C-AJFE-SU-027 under the supervision of Dr. Anna Oldani. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the FAA.

References

[1] M. Colket, J. Heyne, M. Rumizen, M. Gupta, T. Edwards, W. M. Roquemore, G. Andac, R. Boehm, J. Lovett, R. Williams, J. Condevaux, D. Turner, N. Rizk, J. Tishkoff, C. Li, J. Moder, D. Friend, and V. Sankaran, "Overview of the National Jet Fuels Combustion Program," *AIAA J.*, vol. 55, no. 4, pp. 1087–1104, 2017.

[2] "Standard test method for cetane number of diesel fuel oil," ASTM International, Tech. Rep. ASTM D613-18a, 2019.

[3] "Standard test method for determination of ignition delay and derived cetane number (DCN) of diesel fuel oils by combustion in a constant volume chamber," ASTM International, Tech. Rep. ASTM D6890-18, 2019.

[4] J. Heyne, B. Rauch, P. Le Clercq, and M. Colket, "Sustainable aviation fuel prescreening tools and procedures," *Fuel*, vol. 290, p. 120004, 2021.

[5] A. D. Ure and S. Dooley, "Numerical models for the prediction of hydrocarbon physical properties: towards the prescreening of synthetic aviation fuels." 260th Natl. Meet. Am. Chem. Soc., 2020.

[6] Z. Yang, S. Kosir, R. Stachler, L. Shafer, C. Anderson, and J. S. Heyne, "A GC × GC Tier α combustor operability prescreening method for sustainable aviation fuel candidates," *Fuel*, vol. 292, p. 120345, 2021.

[7] J. B. Cooper, C. M. Larkin, J. Schmitigal, R. E. Morris, and M. F. Abdelkader, "Rapid analysis of jet fuel using a handheld near-infrared (NIR) analyzer," *Appl. Spectrosc.*, vol. 65, no. 2, pp. 187–192, 2011.

[8] Y. Wang, Y. Cao, W. Wei, D. F. Davidson, and R. K. Hanson, "A new method of estimating derived cetane number for hydrocarbon fuels," *Fuel*, vol. 241, pp. 319–326, 2019.

[9] E. Al Ibrahim and A. Farooq, "Prediction of the derived cetane number and carbon/hydrogen ratio from infrared spectroscopic data," *Energy Fuels*, vol. 35, no. 9, pp. 8141–8152, 2021.

[10] V. Boddapati, A. M. Ferris, and R. K. Hanson, "On the use of extended-wavelength FTIR spectra for the prediction of combustion properties of jet fuels and their constituent species," *Proc. Comb. Inst.*, 2022.

[11] S. W. Sharpe, T. J. Johnson, R. L. Sams, P. M. Chu, G. C. Rhoderick, and P. A. Johnson, "Gas-phase databases for quantitative infrared spectroscopy," *Appl. Spectrosc.*, vol. 58, no. 12, pp. 1452–1461, 2004.

[12] A. E. Klingbeil, J. B. Jeffries, and R. K. Hanson, "Temperature-dependent mid-IR absorption spectra of gaseous hydrocarbons," *J. Quant. Spectrosc. Radiat. Transf.*, vol. 107, no. 3, pp. 407–420, 2007.

[13] J. Yanowitz, M. A. Ratcliff, R. L. McCormick, J. D. Taylor, and M. J. Murphy, "Compendium of experimental cetane numbers," National Renewable Energy Laboratory, Tech. Rep. NREL/TP-5400-67585, 2017.

[14] Y. Wang, Y. Ding, W. Wei, Y. Cao, D. F. Davidson, and R. K. Hanson, "On estimating physical and chemical properties of hydrocarbon fuels using mid-infrared FTIR spectra and regularized linear models," *Fuel*, vol. 255, p. 115715, 2019.

[15] A. G. Abdul Jameel, N. Naser, A.-H. Emwas, S. Dooley, and S. M. Sarathy, "Predicting fuel ignition quality using 1H NMR spectroscopy and multiple linear regression," *Energy Fuels*, vol. 30, no. 11, pp. 9819–9835, 2016.

[16] H. Drucker, C. J. Surges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in *Advances in Neural Information Processing Systems*, vol. 1, 1997, pp. 155–161.

[17] A. B. Smola and B. Schölkopf, "A tutorial on support vector regression," *Stat. Comput.*, vol. 14, pp. 199–222, 2004.