Neural network-based prediction of liquid-phase diffusion coefficient to model fuel-oil dilution on engine cylinder walls

Published Version:
Mariani V., Pulga L., Bianchi G.M., Cazzoli G., Falfari S. (2020). Neural network-based prediction of liquid-phase diffusion coefficient to model fuel-oil dilution on engine cylinder walls. SAE INTERNATIONAL JOURNAL OF ENGINES, 13(5), 649-664 [10.4271/03-13-05-0041].

Availability:
This version is available at: https://hdl.handle.net/11585/780849 since: 2020-11-17

Published:

DOI: http://doi.org/10.4271/03-13-05-0041

0XX-01-XXXX

Neural Network Based Prediction of Liquid Phase Diffusion Coefficient to Model Fuel-Oil Dilution on Engine Cylinder Walls

Author, co-author (Do NOT enter this information. It will be pulled from participant tab in MyTechZone)

Affiliation (Do NOT enter this information. It will be pulled from participant tab in MyTechZone)

Abstract

Nowadays the role played by passenger vehicles on the greenhouse effect is worth heavy. In order to slowdown both the global warming and the fossil fuels wasting, the design of high efficiency engines is compulsory. Downsized Turbocharged Gasoline Direct Injection engines comply with both high efficiency and power demand requirements. Nevertheless, Direct Injection inside downsized chambers may result in the fuel wall impingement depending on the operating conditions. The impact of fuel on the cylinder liner leads to the mixing between the fuel and the lubricant oil on the cylinder wall. When the piston moves, the piston top ring scraps the non-evaporated fuel-oil mixture. Then, the scraped fuel-oil mixture may be scattered into the combustion chamber, becoming source of diffusive flames in all conditions and abnormal combustions known as Low Speed Pre-Ignitions at the highest loads. In order to analyse these phenomena, accurate predictions of the liquid phase diffusion between fuel and oil are needed. Currently, no experimental data are available for the diffusion between fuel and oil, and common correlations are characterized by high inaccuracy (errors around 20-40% are reported). In this work, a Deep Neural Network methodology was developed and validated against engine-like fluids. Furthermore, the diffusion coefficients for different gasoline surrogates/SAE oils are provided and the effect of gasoline-ethanol blending is discussed.

Introduction

The further tightening of the tailpipe particulate emission limits expected beyond those by Euro 6d-temp and the application of new test cycles (WLTC and RDE) bring new challenges to the design of the engine combustion systems. In the last years, the efficiency and environmental impact of gasoline-powered engines have been greatly improved with the application of downsizing together with other technological solutions like Direct Injection (DI), Miller/Atkinson cam, advanced combustion systems (SPCCI, by Mazda) etc. One of the main drawbacks of the joint use of DI and downsizing is the liquid fuel wall impingement, with emphasis on the cylinder liner and the piston crown. Different technological solutions have been applied to promote faster fuel evaporation and lower spray momentum: injector nozzle shape and drilling; improved spray targeting; multiple injections strategies; ever higher injection pressure (currently 350 bar are available and 500 bar are expected in few years). Nevertheless, the liquid fuel droplet-wall impingement remains a concern. It must be underlined that also Port Fuel Injection (PFI) and Direct Port Injection (DPI) configurations are affected by liquid fuel wall impingement. In both the configurations, the inflow air momentum strips-off the fuel liquid film formed on the intake port walls or on the intake valve seat crevice. Then, the stripped fuel droplets convecting against the cylinder liner or the piston crown realize the wall impingement.

Focusing on the wall impingement against the cylinder liner, it should be remembered that the cylinder wall is wetted by a lubricant film of few micrometres of height, which allows the upward and downward piston motion within the cylinder. The impact of high kinetics fuel droplets against the oil layer on the cylinder wall, may result in the strip-off of liquid droplets, whose composition (fuel only, oil only or fuel-oil mixture) depends on the impact regime [1, 2]. In the current Turbocharged Gasoline Direct Injection (TGDI) engines, due to the high kinetics of the fuel droplets (high injection pressure) and the wet state of the cylinder wall, the *splash* and the *spread* regimes are the most likely fuel impingement regimes [3]. The *spread* regime leads to the fuel deposition, whilst the *splash* regime leads to both the fuel deposition and the scattering of post-impingement liquid droplets. The fuel deposited due to *spread* or *splash*, dilutes with the lubricant oil layer, resulting in a fuel-oil mixture. The dilution between the fuel and the oil causes the degradation of the oil properties, with emphasis on the viscosity, since the own typical values of lubricant oils (≈ 10^2 mPa·s) and fuels (≈ 0.5 mPa·s) highly differ from each other’s, with concerns on the engine reliability. The deposited fuel is added to the oil layer, leading to the local thickening of the wall film on the cylinder liner. When the piston reaches the thicker part of the wall film during the compression stroke, the fuel-oil mixture is scraped into the first piston land crevice. This mixture may be scattered into the combustion chamber during the piston retraction due to inertia forces. The scattered liquid, comprising both fuel and lubricant oil, may cause Pre-Ignition (PI) phenomena or diffusive flames (soot formation) depending on the engine load.

Several Authors reported evidences of the relationship between the presence of flying lubricant oil droplets in the combustion chamber and abnormal combustion events known as Low Speed Pre-Ignition (LSPI). The LSPI may occur in modern TGDI engines when accelerating at low speed and high loads, likely resulting in severe engine damages (spark plug wear, cracked piston). Dahnz et al. [4] conducted both experimental and numerical campaigns with an optical accessible engine. According to the Authors, likely the presence of a second phase (liquid or solid particles) in the combustion chamber is responsible of the LSPI. Amman et al. [5] conducted experiments to test the influence of different engine parameters on the LSPI occurrence frequency such as air/fuel ratio, coolant temperature, fuel enrichment, etc. In [5], the measurements at the exhaust port suggested that during the LSPI events, the gas mixture was enriched inconsistently with the operating conditions. Therefore, the Authors assumed that an extra source of hydrocarbons (HCs) accumulation and release must be present in the combustion chamber. In [6] the same Authors tested difference piston crevice configurations, discovering a relationship between the piston crevice volume and the LSPI occurrence frequency. As a result, the Authors supposed that the piston crevice was the HCs source under investigation. Welling et al. [7] tested a downsized gasoline engine and reported that the PI events might be triggered by the presence of foreign low ignition-delay spots in the combustion chamber. Since the lubricant oil is characterized by a larger chemical reactivity with respect to fuel (lower ignition-delay and ignition temperature), its detachment and transport into the combustion chamber was considered the responsible of those low ignition-delay spots. Furthermore, it must be remembered that the lubricants mainly made by HCs. Therefore, flying lubricant droplets in the combustion chamber, promoted by the piston-induced lubricant film strip-off, may burn individually according to a diffusive combustion, which is a recognized source of soot formation. In a review paper [8] Raza et al. collected several findings on the analysis of the particulate at the tailpipe of TGDI engines. The work reports the significant contribution of the lubricant oil to the overall Particulate Mass (PM) and Particulate Number (PN). Thus, it is reasonable to believe that the lubricant oil detachment and transport in the combustion chamber and the soot emission are correlated.

Since the fuel-oil dilution affects not only the engine lifetime (LSPI, parts wear) but also the advances in engine downsizing and the design of low-impact combustion systems, a deep understanding of the dilution process between fuel and oil is needed. In order to avoid the high time and cost requirements of experimental investigations reproducing LSPI conditions, the numerical modelling is attractive to simulate the diffusion between the fuel and the oil lubricant under engine-like conditions. The experimental measure of the diffusion coefficient is known to be very annoying and expensive in terms of care, time and cost spent. This is true especially for mixtures comprising oils, due to the high viscosity and opacity of these fluids. At present, few Authors have reported models of fuel-oil dilution under engine-like conditions. In these models, the 2nd Fick's law, which needs the diffusion coefficient between fluids, is the standard approach to model the mass transport by diffusion. Yu and Min [9] implemented a diffusion model between the fuel and the oil film by considering two representative mono-component liquids assuming a uniform constant temperature along the thickness. In [9] the diffusion coefficient was initially estimated with the correlation of Hayduk and Minhas [10] for fuel-oil and vice versa at infinite dilution, i.e. solute molar concentrations < 10%. Then, the Authors used that estimations to calculate the diffusion coefficient depending on the solvent and the solute molar concentrations. Zhang et al. [11] improved the model of Yu and Min by implementing additional features such as the multi-component liquid approach and the temperature dependence, solved with the Fourier equation. In spite of Yu and Min that assumed a concentrated fuel-oil mixture, the Zhang et al. considered the fuel-oil mixture as infinite diluted and estimated the diffusion coefficient with the classical Wilke and Chang correlation [12]. It must be considered that the results reported in [9, 11] might be not reliable because the correlations adopted to estimate the diffusion coefficients are affected by significant errors. Siddiqi and Lucas [13] compared the results of different correlations (including the ones by Hayduk and Minhas and Wilke and Chang) with several hundred experimental data. In [13] the Authors reported mean absolute errors from 13% to 20% for organic solutions and from 20% to 35% for aqueous solutions. However, the modelling of the diffusion coefficient remains the most viable method to estimate and predict this key transport property. Based on this background, increasing the accuracy of the methods that model the diffusion coefficients, is a step forward in the analysis and the prediction of phenomena of engine interest.

In the last decade, thanks to the advances in computing power, Machine Learning (ML) and Neural Networks (NNs) techniques have proven to be powerful tools for the prediction of the fluid properties. The liquid-liquid equilibrium [14] and the vapour-liquid equilibrium [15], the thermal diffusivity [16], the fuel laminar flame speed [17, 18] and the diffusion coefficient itself [19], have been yet successfully predicted with those techniques. The present work deals with the implementation of different ML and NN methodologies to predict the diffusion coefficient at infinite dilution. The developed methodologies are based on an extensive database created after a deep literature review and comprising a large number of HCs, accounting for some components of real gasolines and lubricant oils, and alcohols, accounting for gasoline-biofuel blends. The methodologies were validated against the experimental diffusion coefficients of different mixtures given by the combination of HCs. The main contribution of this work is the implementation of the *hybrid* methodologies, where efforts were made to integrate in a physical manner the predictions by ML algorithms and NNs with the predictions by traditional correlations. After the validation step, the methodology called *hybrid W-C*, which integrates a deep NN with the classical Wilke and Chang correlation, showed a significant gain in accuracy, reliability and interpretability with respect to both traditional correlations and stand-alone ML algorithms and NNs, which are the common practice in the current state of the art. Thus, the hybrid W-C methodology was applied to predict the diffusion coefficient of different binary combinations of gasoline surrogates and SAE lubricant oils.

Methodology

This work deals with the development of a numerical methodology to predict the liquid phase diffusion coefficient of binary dilute solutions. Currently, the most common approach to estimate the diffusion coefficient, is the use of empirical and semi-empirical correlations that were developed decades ago. Some remarkable correlations are the Wilke and Chang correlation [12] and the Siddiqi and Lucas correlation [13]. The former is a milestone that inspired a number of investigations and it is strongly based on the classical Einstein-Stokes equation. The latter is one of the most recent steps forward in the topic, and the one with the better agreement with the experimental data. When one investigates the mass transport by diffusion between species, the most common assumption is that the diffusion rate depends on: i) the frictional resistance between molecules; ii) the intermolecular forces. The frictional resistance was commonly represented with the solvent dynamic viscosity, whilst the representation of the intermolecular forces was the target of several interpretations. Based on the comparison with the experimental data, Wilke and Chang adjusted the Einstein-Stokes equation by adding the effective solvent molecular weight as a measure of the intermolecular forces. The effective molecular weight was defined as the product of the solvent molecular weight with the so called *association factor*, i.e. a multiplicative coefficient representative of the bonds strength in the liquid. Siddiqi and Lucas assumed that calculating the molar volumes of both the solvent and the solute at their own Normal Boiling Point (@NBP) was sufficient to take into account the intermolecular forces. For sake of clarity, the correlations by Wilke and Chang and Siddiqi and Lucas are reported respectively in Eq. (1) (generally valid) and in Eq. (2-3) (respectively valid for organic and aqueous solutions). In Eq. (1-3) *T* is the liquid temperature, μ is the liquid dynamic viscosity, *M* is the molecular weight, V_b is the molar volume @NBP, *A* is the association factor and *K* is a multiplicative constant. The subscripts 1 is associated with the solvent, whilst the subscript 2 is associated with the solute. The quantities with no subscripts (*T*, μ) are associated with the mixture (solvent + solute).

$$D_{12} = K\left[\frac{T}{\mu}\,\frac{(A_1\,M_1)^{0.5}}{V_{b2}^{0.6}}\right], K = 7.40 \text{ x } 1\text{E} - 12 \tag{1}$$

$$D_{12} = K\left[\frac{T}{\mu^{0.907}}\,\frac{V_{b1}^{0.265}}{V_{b2}^{0.45}}\right], K = 9.89 \text{ x } 1\text{E} - 12 \tag{2}$$

$$D_{12} = K\left[\frac{T}{\mu^{1.026}}\,\frac{1}{V_{b2}^{0.5473}}\right], K = 2.98 \text{ x } 1\text{E} - 11 \tag{3}$$

Fig. (1) shows the Absolute Relative Error (ARE) committed with the Wilke and Chang correlation and the Siddiqi and Lucas correlation when applied to different solvent/solute combinations. The upper part of the figure shows mixtures that were commonly characterized in the experimental measure of the diffusion coefficients. The lower part of the figure shows HCs mixtures where the solvent is a light HC (number of carbons $C < 10$) and the solute is a heavy HC ($C > 10$). These solvent/solute combinations are the closer to fuel-oil mixtures, where the solvent (gasoline) and the solute (lubricant oil) are blends of HCs that has an average number of carbons respectively between *C*7-*C*8 and *C*30-*C*35 [20]. In Fig. (1) is visible that the error committed on the fluids in the upper part of the figure is small (< 10%), whilst the error committed on the light/heavy HCs mixtures is high (> 10% for Eq. (2) and > 20% for Eq. (1)). In the lower part of Fig. (1) is visible that the heavier are the solvent and the solute molecules, the higher are the errors committed by the correlations. It needs to be considered that n-heptane (*C*7) and n-octane (*C*8) are similar to real gasolines in terms of weight and average carbons number, whilst n-tetradecane (*C*14) and n-hexadecane (*C*16) are often the lighter part of the lubricants. Therefore, since *C*30-35 and *C*40-50 are representative of the average and the maximum carbons number in the lubricant oils, a further increase of the error committed by those correlations is expected for realistic engine mixtures. These errors are in contrast with the ever higher need of accuracy and detail in engine modelling. This work approaches the prediction of the diffusion coefficient by using different ML algorithms and NNs. Since these techniques require a large number of reliable data in order to be trained and experienced properly, an extensive database was created.

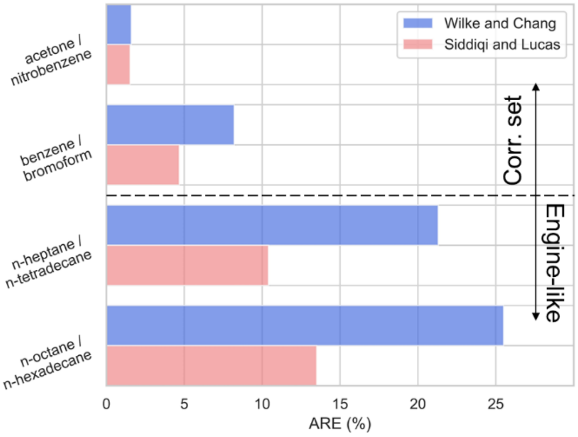

Figure 1. Absolute Relative Error with respect to experimental data for the Wilke and Chang (Eq. (1)) and the Siddiqi and Lucas (Eq. (2)) correlations

Database creation

Consider the TGDI engine conditions at low engine/piston speed and wall film close to the TDC. These conditions result in maximum oil dilution with gasoline on the cylinder liner. The time available for mass diffusion between fuel and oil in under these conditions in a stroke before piston arrival is around 10 milliseconds. Considering this period as the time interval to integrate the 2nd Fick's law, and assuming the typical value of 1E-9 m^2/s for the diffusion coefficient, concentration changes below the 10% are expected. The aforementioned experimental evidences suggest that this degree of dilution is sufficient to affect the engine operations, however they result too low to assume the fuel-oil mixture as concentrated. Hence, in the present work the fuel-oil mixture was considered a dilute solution. Therefore, the experimental liquid phase diffusion coefficients of binary mixtures at infinite dilution were collected to create the database. Moreover, the database creation was addressed towards liquids that are representative of the components of real gasolines (e.g. cyclohexane, toluene, n-heptane, n-octane, i-octane, n-decane, n-dodecane, ethanol, methanol, buthanol) and oils (e.g. oleic acid, n-tetradecane, n-hexadecane, kerosene). Following the abovementioned criteria, about 250 mixtures given by the combination of 72 different liquids were collected from experimental findings in literature [21-47]. According to some literature findings [12, 13, 48] and considerations of the present Authors, the fluid properties that mainly affect the diffusion coefficient were selected. Then, these properties were provided as *input features* to the ML algorithms and the NNs. As a result, the dynamic viscosity (μ), the molecular weight (*M*), the molar volume (*V*), the density (ρ), the latent heat of vaporization (L_V), the number of carbons (#*C*), hydrogens (#*H*) and oxygens (#*O*) were chosen. For each liquid, these properties were collected as follows: i) the density and the dynamic viscosity from literature experimental data at ambient conditions (298 K, 1 atm); ii) the latent heat of vaporization at the saturation point from experimental data if available, otherwise by means of correlations [38, 49]; iii) the molar volume at ambient conditions and at NBP respectively by means of Eq. (4) and Eq. (5). In Eq. (5) The liquid density @NBP was estimated according to the correlations reported in [38] by replacing the generic temperature with the normal boiling temperature at zero vapour fraction.

$$V = \frac{M}{\rho} \tag{4}$$

$$V_b = \frac{M}{\rho_{@NBP}} \tag{5}$$

Machine Learning methodologies

In this work, three Machine Learning methodologies to predict the diffusion coefficient, which are deeply described in the following section, have been developed and compared: a) pure regression; b) hybrid Wilke-Chang (hybrid W-C); c) hybrid Siddiqui-Lucas (hybrid S-L). For each methodology, several regression algorithms and NNs have been tested while maintaining the same optimization workflow of the tuning parameters (*hyperparameters*). The implementation of these algorithms is based on open-source software libraries developed for Python 3.7.4: Scikit-learn [50] for data pre-processing and classical ML models, Keras [51] and Tensorflow [52] for the NNs.

Considering the nonlinear dependencies between the fluid properties selected as input features (μ, *M*, *V*, ρ, *Lv*, *#C*, *#H*, *#O*) and the target variable, several new features have been generated. These new features, called *derived features*, were obtained as the ratio between the solvent and the solute properties and by applying exponentials to the physical properties. The generation of derived feature*s*, together with the limited size of the database in comparison with those recommended for ML methods [53], increases the risk of overfitting, i.e. generate a regression algorithm with excellent performance on the training set, but not able to generalize to new data. In order to limit this risk, only the most meaningful features were selected based on a rank of mutual information. Fig. (2) shows the mutual information rank for the ten most relevant among features and derived features. The mutual information between two discrete sets (X and Y) is a measure of the dependence of a variable to the other that does not rely on their covariance. Thus, it allows for nonlinear dependencies to be accounted for. Its value is calculated as in Eq. (6), where $p_{(X,Y)}$ is the joint probability mass function of X and Y, while p_X and p_Y are the marginal probability mass functions of X and Y.

$$I(X,Y) = \sum_{x\in X} \sum_{y\in Y} p_{(X,Y)}(x,y)\, log\left(\frac{p_{(X,Y)}(x,y)}{p_X(x)\; p_Y(y)}\right) \tag{6}$$

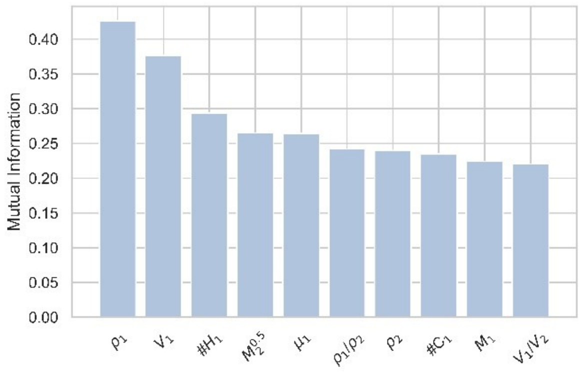

Figure 2. Example of features and derived features ranked with decreasing mutual information

After applying the mutual information analysis of each original feature and derived feature with the target, only the features with a relative importance greater than 0.05 were considered (Table (A1)). This allows the training step of the ML algorithm to rely only on the most meaningful features and reduces the risk of "memorizing" the input data.

After the mutual information rank, the optimization of the regression algorithms has been performed by applying a repeated k-fold approach [54] combined with an extensive grid search among the hyperparameters of the models. For this application, repeated k-fold consisted in iteratively splitting the database into train and test set, normalizing the features with mean value set to 0 and standard deviation set to 1. Then, the algorithm is trained on the train set and the accuracy of the model is evaluated on the test set as the mean of all the repetitions. This approach is required to avoid defining hyperparameters and NN architectures based on observations that depend on a single dataset split. Among the tested regression algorithms, AdaBoost [55] and Feed Forward Neural Networks (FFNN) [54] have performed best in terms of accuracy. The accuracy goal is checked by comparing the coefficient of determination (R^2), Eq. (7), and the Mean Absolute Relative Error (*MARE*), Eq. (8), where y_i and $\bar{y}$ are the real targets and their mean value, $\tilde{y}_i$ is the predicted value.

$$R^2 = 1 - \frac{\sum_i (y_i - \tilde{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{7}$$

$$MARE = \sum_i \frac{|y_i - \tilde{y}_i|}{y_i} \tag{8}$$

Focusing on AdaBoost, Zhu [55] and Freund [56] were among the first to introduce the idea of boosting as a strategy to enhance the performance of several simple regression algorithms by computing a weighted average of their outputs to obtain the final prediction. The algorithm of AdaBoost is based on the successive training of base learners on a modified version of the training set, where more weight is given to the samples that were not predicted accurately by the previous models. The most relevant parameters of this algorithm are the number, type and characteristics of base learners and the learning rate, which controls the contribution of each regressor to the final output.

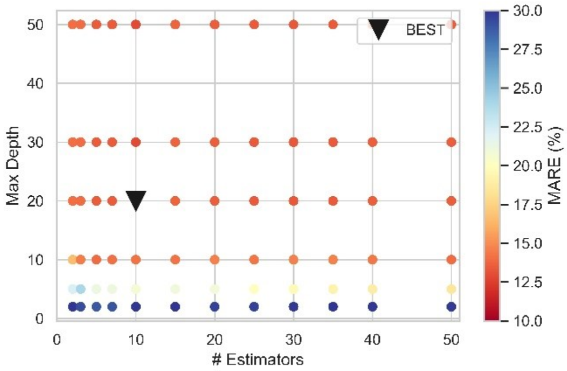

Figure 3. Mean Absolute Relative Error map and example of the optimized point (black marker)

To limit the search of the parameters, the base learner has been set to a decision tree regressor [52], while the number of estimators, the maximum depth of the tree and the learning rate have been optimized. It must be underlined that in this work, adding more estimators would have not improved the performance of the algorithm on both train and test sets. On the other hand, the depth of each regression tree is a parameter that allows to better fit the train set and therefore its value must be carefully controlled to avoid overfitting. This can be noticed by the increase of *MARE* for the largest values of the ordinate in Fig. (3).

A FFNN applied to a regression task is a multi-layer network of simple elements, called *neurons*, that receive the input features on one end of the network, and provide a predicted value on the other end, without any recurrence or matrix manipulation step [53]. Neurons are organized in layers (of any width) connected with each other from the input layer (of the same size as the number of features), to the output layer. The information is transferred from one layer to the next one via connections among the neurons of variable weights, which are optimized during the training step, and then modified by applying a transformation function (usually hyperbolic tangent or rectified linear unit [56]). The parameters that might affect the performance of NNs the most are the topology of the network (number of neurons and distribution on different layers) and the activation function. The optimization algorithm employed for the weights update has been kept constant (Adam [56]), as well as the kernel initializers (normal distribution with std = 0.05 and mean value = 0). The number of epochs (number of times that the full dataset is used to update the weights of the neurons [53]) allowed during the training step has been set to 1E5. This stop criterion was integrated with a control strategy that would stop the training of the dataset if the performance on the training set would not improve after 20 consecutive iterations. Considering the limited size of the train set, the neurons weights would tend to "memorize" the training points and to overfit the data. In order to avoid this risk, each NN has been integrated with an additional input layer where Gaussian normal noise is added to the normalized features. Moreover, one of the internal layers has a kernel L2 regularization to improve the robustness of the prediction. The standard deviation of the Gaussian noise and the lambda value [57] of the L2 regularizer have been added to the optimization grid search.

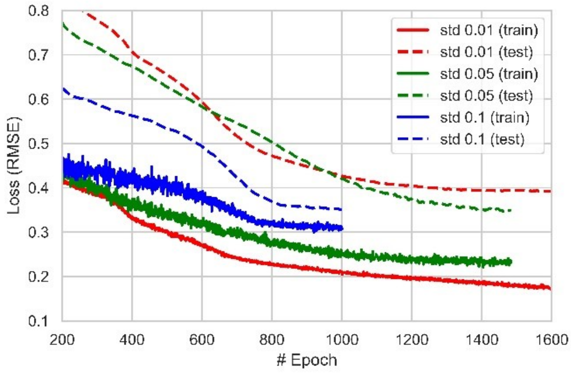

Figure 4. Root Mean Square Error for different amplitude of the Gaussian noise

In Fig. (4), the effect of different values of std of the Gaussian noise on the training performance of a NN is reported as a function of the epochs and the loss (Root Mean Squared Error (RMSE)). It can be noticed that the performance of the algorithm on the training set increases with the number of epochs until a minimum asymptotic value, while the performance of the same model on the test set suffers from increasing overfitting. The addition of increasing Gaussian noise to the input features reduces the distance from the test to the train loss, even if the performance on the train decreases. Considering the early stopping strategy after a given number of successive steps, it can be noticed that as the noise increases, the optimum value is reached after a smaller number of time steps. Several structures of NNs have been tested during the optimization step, with a different number of layers, neurons per layer and activation functions. For sake of simplicity, the illustration of the architecture of the optimized NNs is avoided, thus, a schematic representation of a deep NN general structure with one output neuron is given in Fig. (5).

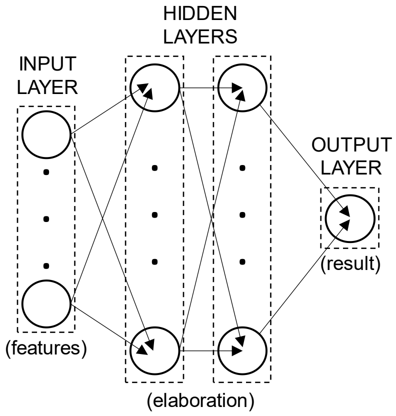

Figure 5. Schematic structure of the implemented deep Neural Networks

Pure regression methodology

The pure regression methodology consisted in the prediction of the diffusion coefficient by directly applying the regression algorithm (AdaBoost or FFNN). Thus, the regression task of the ML technique, called θ, is the diffusion coefficient itself ($D_{12} = \theta$). In order to avoid the risk of fading gradients for deeper NNs [52], given the order of magnitude of the diffusion coefficient in SI units, the target value has been scaled by 1E9.

Hybrid methodology

The two hybrid methodologies (hybrid W-C and hybrid S-L) were developed by implementing a three-steps prediction using both the Machine Learning technique (AdaBoost or FFNN) and an empirical correlation (the Wilke and Chang correlation in the hybrid W-C and the Siddiqi and Lucas correlation in the hybrid S-L). The three steps can be resumed as follows: i) in the first step the ML technique, given the top ranked input features, is trained to return a corrective factor that is the regression task (θ) of the hybrid methodologies; ii) in the second step the empirical correlation is applied in its standard form (regardless to the mutual information); iii) the diffusion coefficient returned by the correlation is multiplied by the correction factor (θ). A scheme of the abovementioned workflow is shown in Fig. (6).

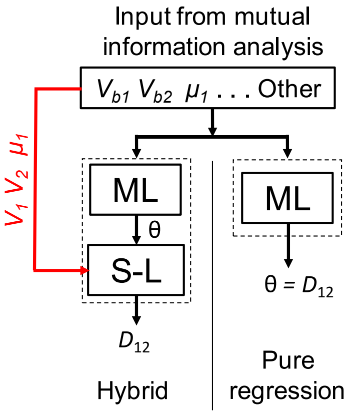
Input from mutual information analysis
V_{b1} V_{b2} μ_1 . . . Other
V_1 V_2 μ_1
ML
θ
S-L
D_{12}
Hybrid
ML
θ = D_{12}
Pure regression

Figure 6. Comparison between the flowchart of the pure regression and the hybrid (hybrid S-L) methodologies

In the hybrid W-C methodology, the Wilke and Chang correlation was rearranged as follows: i) the solvent molecular weight was replaced with the molar volume according to Eq. (4); ii) the molar volumes at ambient conditions were used instead of the ones @NBP; iii) the association factor was removed; iv) under the assumption of dilute solution, the temperature and the dynamic viscosity were associated with the solvent instead of the mixture. Due to points ii) and iii), the rearranged correlation lacks of a measure of the intermolecular forces. The original idea of Wilke and Chang to take into account the intermolecular forces by introducing a correction, i.e. the association factor, was maintained. To this aim, the θ correction factor predicted by the ML step adjusts the solvent-solute molar volume ratio at ambient conditions, which is increased or decreased depending on the bonds strength in the two liquids. In Eq. (9) the final formulation of the diffusion coefficient predicted with the hybrid W-C methodology is shown. Fig. (7) shows the θ correction factor predicted for the train set against the carbons number of the solvent. The zero and the low carbon zones, which include solvents with strong bonds (H bonds) such as water and alcohols, are characterized by the highest correction factors. Then, the higher is the carbons number, i.e. the weaker are the bonds, the lower is the correction factor. Since the correction factor is not normally distributed in the dataset, a log transformation is performed on the target values before training and reversed after the regression step in order to improve the performance of the regression algorithms.

$$D_{12} = K\left[\frac{T_1}{\mu_1}\rho_1{}^{0.5}\left(\theta\frac{V_1{}^{0.5}}{V_2^{0.6}}\right)\right] \qquad (9)$$

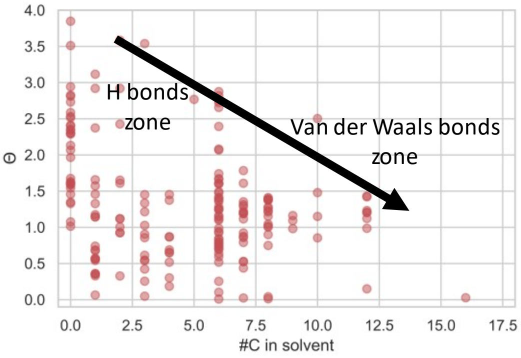

Figure 7. Correction factors predicted by the hybrid W-C against the carbon number of solvents

In the hybrid S-L methodology, the Siddiqi and Lucas correlation was fully maintained except for the application of the same point iv) as for the hybrid W-C methodology. In the hybrid S-L methodology, the ML step predicts a correction factor θ which is a pure multiplicative coefficient as reported in the scheme in Fig. (6) and in Eq. (10) (organic solutions).

$$D_{12} = \theta\left[K\left(\frac{T_1}{\mu_1^{0.907}}\frac{V_{b1}^{0.265}}{V_{b2}^{0.45}}\right)\right] \qquad (10)$$

Optimized methodologies

The optimized regression algorithm AdaBoost is summarized in Table (1) for the three implemented methodologies. The FFNN optimization led to three different architectures. The optimized FFNN is composed by 2 hidden layers of 12 neurons each in pure regression mode, 2 hidden layers of 30 neurons each in hybrid S-L mode, and 3 hidden layers of 20 neurons each in hybrid W-C mode. In pure regression and hybrid S-L modes, the activation functions are the tanh for the first hidden layer and the ReLu for the second. In hybrid W-C mode, the activation functions are the tanh for the first hidden layer and the ReLu function for the other two. The Gaussian noise std is set to 0.02 in pure regression and hybrid S-L modes and to 0.01 in hybrid W-C mode. The lambda for L2 regularizer is set to 0.01 in all the three methodologies. Table (2) provides a brief view of the abovementioned optimized NN architectures.

Table 1. Optimized AdaBoost parameters for the three methodologies

	Learning rate	**Base estimators**	**Max depth**
Pure regression	0.1	30	10
Hybrid W-C	0.1	20	15
Hybrid S-L	0.1	35	18

Table 2. Optimized Neural Network structures for the three methodologies

	Hidden layers	Neurons/layer
Pure regression	2	12
Hybrid W-C	3	20
Hybrid S-L	2	30

Results

Results on the train set

Table (3) reports the performance of the optimized AdaBoost and NN on the train test after repeated k-fold cross validation (10 repetitions on each of the 5-fold splits) for the three implemented methodologies. As visible in Table (3), for all the methodologies the NN shows higher R^2 mean values (especially in pure regression mode) and R^2 standard deviations that are half of the ones reported for AdaBoost. According to this, the NN resulted the winner ML technique. Thus, in the following sections the results of the pure regression, the hybrid W-C and the hybrid S-L, are presented only with the NNs.

Table 3. Prediction performance for AdaBoost and Neural Networks on the train set

		Mean R^2	Std R^2
Pure regression	NN	0.95	0.22
	AdaBoost	0.90	0.40
Hybrid W-C	NN	0.96	0.15
	AdaBoost	0.95	0.25
Hybrid S-L	NN	0.95	0.12
	AdaBoost	0.92	0.34

Results on the full dataset

In order to compare the performance of the three implemented methodologies with particular focus on HCs mixtures, a targeted split of the dataset has been performed. The split between train and test set has been performed targeting the representability of the two sets of features in the range of application of HCs mixtures, and the similarity between the distribution of the target variables in both sets. The validation set was created including 15 different combinations of butanol, n-hexane, n-heptane, n-octane, n-decane, n-dodecane, n-tetradecane, n-hexadecane, kerosene and oleic acid. These fluids were selected directly by the present Authors, since they were the more similar ones to the fluids of interest, i.e. gasolines and lubricant oils, with experimental characterization data of their physical properties available.

The optimized NNs for each methodology have been trained on the train set and the results are reported in Table (4) in terms of R^2 and *MARE* for the train set, the test set and the validation set. Despite similar performances on the train and test set, the hybrid W-C methodology has achieved the better results ($R^2 = 0.98$, $MARE \approx 6\%$) on the validation set. The comparison in Table (4) shows that the performance of the regression task is improved by the integration of the NN within the correlation. Moreover, the potential of the hybrid methodologies is the interpretability of the most relevant dependencies that are not assigned to the regression task.

Table 4. R^2 and Mean Absolute Relative Error for three different Neural Network methodologies on train, test and validation sets

	Train		Test		Validation	
	R^2	MARE (%)	R^2	MARE (%)	R^2	MARE (%)
Pure regression	0.99	11.3	0.60	26.6	0.94	21.4
Hybrid W-C	0.99	6.3	0.72	14.3	0.98	6.2
Hybrid S-L	0.99	6.7	0.68	15.8	0.96	10.2

Fig. (8) shows the deviation of the diffusion coefficient predicted with the hybrid W-C methodology by the experimental values for the train and the test set. The figure highlights a good agreement with the experiments, in particular for values lower than one, where real gasolines and oils are expected based on the performed literature review [32-35]. The ARE shows the cumulative distributions reported in Fig. (9) for the hybrid W-C methodology and the two benchmark correlations (Eq. (1), Eq. (2,3)). A general reduction of the mean error that can be attributed to the train set is observed. However, the gain in reliability when one adopts the proposed methodology is underlined by the fact that the fraction of points below the 5% error is about the 70% of the database, while the fraction of points below 15% error is almost the full database (about 90%).

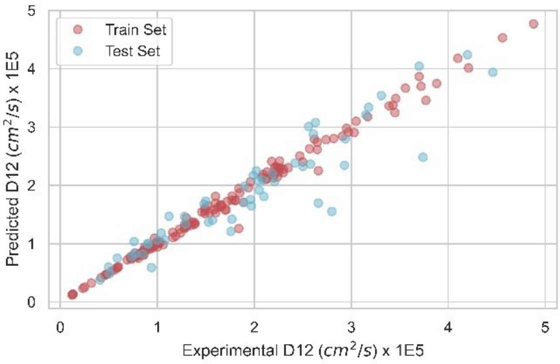

Figure.8 Comparison between experimental and predicted (hybrid W-C) diffusion coefficient

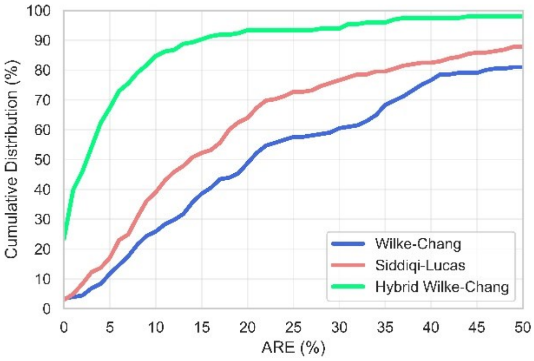

Figure 9. Absolute Relative Error cumulative distribution in predicting the database for two correlations and the hybrid W-C methodology

Diffusion coefficient for gasoline-lubricant mixtures

Since the work aims to predict the diffusion rate between gasolines and lubricant oils for engine modelling, five gasoline surrogates (one mono-component and four multi-component) and two SAE oils (one single-grade and one multi-grade) were considered to create different fuel-oil (solvent-solute) combinations (Table (A2)). For the multi-component liquids, the properties that are required by the hybrid W-C methodology were calculated as follow: i) the mass-weighted average (Eq. (11)) for the density, the molecular weight, the molar volume and the latent heat of vaporization; ii) the Grunberg-Nissan law [61] for dynamic viscosity (Eq. (12)).

$$\phi_{mix} = \sum_{i=1}^{Ncomp} w_i \, \phi_i \tag{11}$$

$$\log(\mu_{mix}) = \sum_{i=1}^{Ncomp} x_i \, log(\mu_i) \tag{12}$$

For sake of comparison, in addition to the hybrid W-C methodology, the predictions of the gasoline surrogate/SAE oil mixtures were performed with the benchmark correlations (Eq. (1,2)). According to the original formulation proposed by Wilke and Chang, in Eq. (1) the association factor for the gasoline surrogates was set to 1, since the HCs are classified as non-associate fluids. It must be noticed that for this calculation the molar volume at ambient conditions (298 K, 1 bar) was used in Eq. (1,2) instead of the molar volume @NBP as that adopted in the original formulation. This choice was taken because both gasolines and the lubricant oils are blends of hundreds of species and their own normal boiling temperature uses to be comprised in wide ranges.

Table 5. Predictions of different gasoline/oil combinations with the hybrid W-C NN methodology, the Wilke and Chang correlation and the Siddiqi and Lucas correlation

Solvent/solute	D_{12} (m^2/s) x 1E9		
	Hybrid W-C	Eq. (1)	Eq. (2)
MIT/SAE 10W30	0.702	1.082	1.260
Ford-Synfuel/SAE10W30	0.622	0.965	1.121
RD587/SAE10W30	0.824	1.254	1.433
RON95/SAE10W30	0.821	1.232	1.402
I-octane/SAE10W30	0.803	1.210	1.370
MIT/SAE30	0.634	0.940	1.134
Ford-Synfuel/SAE30	0.565	0.839	1.009
RD587/SAE30	0.751	1.089	1.290
RON95/SAE30	0.754	1.071	1.262
I-octane/SAE30	0.739	1.052	1.233

The diffusion coefficients predicted with the hybrid W-C methodology and the original Wilke and Chang and Siddiqi and Lucas correlations for mixtures of gasoline surrogates and oils are reported in Table (5). In general, as visible in Table (5), the single-grade SAE 30 oil is less diffusive in gasolines than the multi-grade SAE 10W30 oil due to its higher weight and molar volume. Two multi-component surrogates, i.e. the RON95 and the RD587, are more diffusive than the mono-component surrogate (i-octane), whilst the other two multi-component surrogates, i.e. MIT and Ford-Synfuel, are less diffusive. This is due to the higher content of the lighter HCs (e.g. $C5$, $C6$, Table (A2)) that are present in the RON95 and the RD587 surrogates, which contributes to reduce the viscosity, the molar volume and the θ correction (higher latent heat of vaporization) of the average liquids.

In order to compare the results in Table (5), consider that the experimental diffusion coefficient of n-dodecane ($C12$) and n-hexadecane ($C16$) in n-octane (C8) reduces from 1.64 to 1.5 x1E-9 m^2/s as the solute increases from $C12$ to $C16$. It must be remembered that while the n-octane is a reasonable representation of the real gasolines thermo-physical properties, the average and the maximum carbons number of lubricant oils are respectively around 30 and 50. At this point, if one assumes to maintain $C8$ as a solvent while using molecules with $C30$-$C50$ as a solutes, a further decrease of the diffusion coefficient to values lower than 1 x1E-9 m^2/s can be expected. The predictions performed by means of Eq. (1,2) for i-octane/SAE 10W30 (Table (5), 5th row, columns 3 and 4) are respectively 1.21 and 1.37 x1E-9 m^2/s. These values seem to be unreasonably close to the experimental diffusion coefficient for n-octane/n-hexadecane (1.5 x1E-9 m^2/s) with respect to the high increase of carbons number of the solute. Considering this abrupt flattening with respect to the solute change, one can assume that the correlations are not able to capture the influence of solvents and solutes having higher carbons number, thus, the behaviour of heavy high viscosity liquids such as HC oils. The predictions performed by means of the hybrid W-C NN are closer to the expectations for solutes with the average carbons number of lubricant oils. Analysing the data reported Table (5), one can see that the average predictions performed with Eq. (1) and Eq. (2) overestimate the diffusion coefficient respectively by the 48.3% and the 73.5% with respect to the hybrid W-C NN. This difference mainly depends on two factors: i) the correlations were based on a limited number of species, being developed over 50 years ago when much fewer data were available, whilst the hybrid W-C NN is based on data that comprise a large number of different species; ii) the results returned by the correlations strongly depend on the fitting coefficients of the molecular weight and the molar volume that the original Authors tuned based on the experimental dataset. The hybrid W-C NN relies on the capability to interpret the experimental data to capture the complex key dependence on the intermolecular forces resulting in the θ correction.

It must be underlined that since these gasoline/oil combinations were created by the Authors for the sake of the analysis, there are no experimental data available to perform a direct validation. Nevertheless, in order to provide a proof of reliability of the predictions returned by the hybrid W-C for the mixtures of gasolines and lubricant oils, a further prediction test was performed. To this aim, the work by Hiss and Cussler [62] was considered as a reference. In [62] the Authors measured the diffusion coefficient at ambient conditions of n-hexane and naphthalene, which played the role of the solute, in different HC oils, which played the role of the solvent. These HC oils were characterized by a molecular weight in the range 209-667 g/mol and by a dynamic viscosity in the range 3-5000 mPa·s. In [62], the experimental diffusion coefficient of n-hexane diluted in HC oils with a dynamic viscosity in the range 50-300 mPa·s, which is representative of the most common SAE oils, are around 0.15-0.03 x1E-9 m^2/s. For sake of illustration, the hybrid W-C methodology was used to predict the diffusion coefficient of n-hexane (solute) in SAE 10W30 and SAE 30 (solvents) and for all the mixtures listed in Table (5) by reversing solvent and solute. In Fig. (10) one can see that both the order of magnitude and the value of the SAE oils/n-hexane combinations are in agreement with the experimental diffusion coefficient of n-hexane diluted in generic HC oils. Since real gasolines show higher carbons number (*#C* between 7 and 8 are common), molar volume and lower latent heat of vaporization than those of n-hexane, if one maintain the same solvent (SAE oil), reduced molar volume ratio, θ correction factor and diffusion coefficients are expected. As visible in Fig. (10), the predicted diffusion coefficients for the reversed mixtures of Table (5) stay below the reference experimental curve of the n-hexane. As a result, the hybrid W-C methodology has shown to capture the diffusive behaviour of these fluids in the case of gasolines diluted in lubricant oils. Thus, one can expect that the proposed methodology performs similarly in the case of lubricant oils diluted in gasolines, which is the concentration ratio of interest for LSPI analysis in engines.

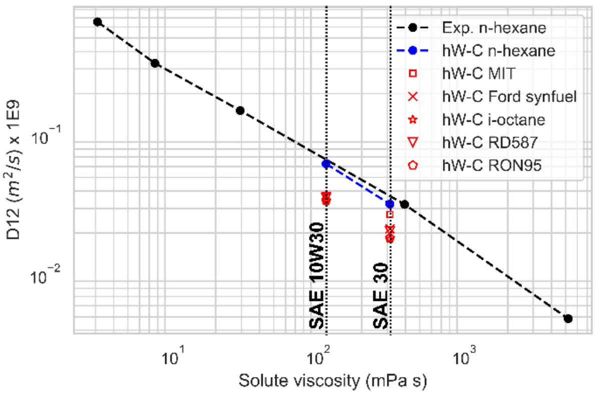

Figure 10. Diffusion coefficient between n-hexane and gasoline surrogates (solutes) in hydrocarbon oils of increasing viscosity (solvent)

Even though the results in Table (5) and Fig. (10) are not directly validated, they were obtained with a new approach in the attempt to provide helpful values to be used in future engine modelling works. Moreover, these results aim to encourage future experiment campaigns of characterization of those fluids, in order to confirm or disprove the present predictions and also to extend the available experimental data to improve the NN methodology.

Fig. (11) shows the predicted effect of blending gasolines (i-octane and RON95 surrogate) with different ethanol volume percentage. Blends with 5%, 10% and 20% of ethanol were considered since they are representative of the most common current gasoline blends with biofuels that can be used without any engine modification. As visible in Fig. (11), regardless to the gasoline composition, the addition of ethanol reduces the diffusion coefficient promoting a slower dilution. The single-grade SAE 30 oil and the multi-grade SAE 10W30 oil are expected to behave similarly with respect to the ethanol addition. The addition of 5%v of ethanol reduces the diffusion coefficients by around the 7%, whilst the addition of 20%v of ethanol reduces the diffusion coefficients by around the 25%.

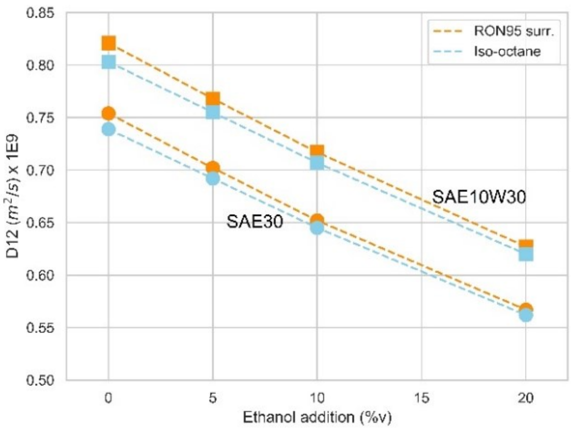

Figure 11. Effect of ethanol addition in gasoline surrogate/oil diffusion

Currently, gasoline blends with ethanol are well-known to promote lower air pollution and greenhouse gases emissions in comparison with fossil fuels. As shown in Fig. (11), a further benefit of the ethanol addition is the reduced diffusion rate of the deposited fuel with the lubricant oils. As a consequence, biofuels blends would be effective to reduce the LSPI frequency of TGDI engines.

Conclusions and future works

The present paper aims to improve the numerical modelling of the fuel-oil dilution process that occurs on the cylinder walls of downsized TGDI engines. To this aim, three different methodologies based on Machine Learning techniques were developed: i) one called pure regression, where a Machine Learning algorithm and a deep Neural Network directly predicts the diffusion coefficient; ii) two called hybrid, where a Machine Learning algorithm and a deep Neural Network are integrated within empirical correlations in a physical manner. The hybrid methodologies are proposed as novel approaches where the Machine Learning technique is used to predict a factor that accounts for the bonds strength of the mixing liquids. This is a complex dependence that the most adopted correlations have proven to be not capable to capture, in particular for fluids of engine interest. After the validation step, the hybrid methodology that integrate a deep Feed-Forward Neural Network within the classical Wilke and Chang correlation (hybrid W-C), has shown to meet the accuracy, simplicity and reliability targets with particular focus on the mixtures of hydrocarbons.

The hybrid W-C methodology was used to predict the diffusion coefficients of different mixtures of gasoline surrogates and SAE lubricant oils. These predictions have shown values around 0.6-0.8 x1E9 m*2*/s, which are consistent with the diffusion coefficients expected for those species based on the observations of experimental data. Furthermore, the behaviour of the diffusion rate with respect to the increasing of the carbons and hydrogens number of both the solute and the solvent, was captured by the proposed methodology. Moreover, the hybrid W-C methodology was used to predict the effect of ethanol addition in gasolines. According to the results, blending common gasolines (pure fossil fuel) with ethanol (biofuel) contribute to strongly slowdown the dilution between the fuel deposited on the cylinder wall and the lubricant oil, likely leading to the reduction of the LSPI frequency.

The hybrid W-C methodology was needed by the present Authors as a key step to develop an in-house One-Dimensional model for the assessment of the LSPI risk under different engine configurations. The accuracy shown by the hybrid W-C methodology, allows the implementation a fast and reliable tool intended to qualitatively support and address the very early stage step of the engine development. This would be performed by comparing the scraped oil mass of two or more configurations under review with different wall film location (i.e. spray pattern), wall temperature, air-fuel ratio, oil species, fuel species, oxygenates addition etc. In this code, the engine variables listed above, would play the role of parameters given by the user in a certain range, resulting in LSPI risk maps.

References

1. Uchida, R., Tanaka, D., Noda, T., Okamoto, S. et al., "Impingement Behavior of Fuel Droplets on Oil Film", *SAE Technical Paper* 2015-01-0913, 2015, https://doi.org/10.4271/2015-01-0913

2. Kassai, M., Torii, K., Shiraishi, T., Noda, T. et al., "Research on the Effect of Lubricant Oil and Fuel Properties on LSPI Occurrence in Boosted S. I. Engines", *SAE Technical Paper* 2016-01-2292, 2016, https://doi.org/10.4271/2016-01-2292

3. Matsuda, T., Senda, J., "Modeling on Spray-Wall Interaction for Direct Gasoline Injection Engines", *Transactions of the Japan Society of Mechanical Engineers Series B*, Vol. 69, No. 688, pp. 2698-2705, 2003, https://doi.org/10.1299/kikaib.69.2698

4. Dahnz, C., Han, K., Spicher, U., Magar, M. et al., "Investigations on Pre-Ignition in Highly Supercharged SI Engines", *SAE Int. J. Engines* 3(1):214-224, 2010, https://doi.org/10.4271/2010-01-0355

5. Amann, M., Mehta, D., and Alger, T., "Engine Operating Condition and Gasoline Fuel Composition Effects on Low-Speed Pre-Ignition in High-Performance Spark Ignited Gasoline Engines", *SAE Int. J. Engines* 4(1):274-285, 2011, https://doi.org/10.4271/2011-01-0342

6. Amann, M., Alger, T., Westmoreland, B., and Rothmaier, A., "The Effects of Piston Crevices and Injection Strategy on Low-Speed Pre-Ignition in Boosted SI Engines", *SAE Int. J. Engines* 5(3):1216-1228, 2012, https://doi.org/10.4271/2012-01-1148

7. Welling, O., Moss, J., Williams, J., and Collings, N., "Measuring the Impact of Engine Oils and Fuels on Low-Speed Pre-Ignition in Downsized Engines", *SAE Int. J. Fuels Lubr.* 7(1):1-8, 2014, https://doi.org/10.4271/2014-01-1219

8. Raza, M., Chen, L., Leach, F., Ding, S., "A Review of Particulate Number (PN) Emissions from Gasoline Direct Injection (GDI) Engines and Their Techniques", *Energies*, Vol. 11, Iss. 6, No. 1417, 2018, https://doi.org/10.3390/en11061417

9. Yu, S., Min, K., "Effects of the Oil and Liquid Fuel Film on Hydrocarbon Emissions in Spark Ignition Engines", *Proceedings of the Institution of Mechanical Engineers*, Vol. 216, Iss. 9, pp. 759-771, 2002, https://doi.org/10.1243%2F09544070260340853

10. Hayduk, W., Minhas, S., "Correlations for Prediction of Molecular Diffusivities in Liquids", *Canadian Journal of Chemical Engineering*, Vol. 60, Iss. 2, pp. 295-299, 1982, https://doi.org/10.1002/cjce.5450600213

11. Zhang, Q., Kalva, V.T., and Tian, T., "Modeling the Evolution of Fuel and Lubricant Interactions on the Liner in Internal Combustion Engines", *SAE Technical Paper* 2018-01-0279, 2018, https://doi.org/10.4271/2018-01-0279

12. Wilke, C. R., Chang, P., "Correlation of Diffusion Coefficients in Dilute Solutions", *AIChe Journal*, Vol. 1, Iss. 2, pp. 264-270, https://doi.org/10.1002/aic.690010222

13. Siddiqi, M. A., Lucas, K., "Correlations for Prediction of Diffusion in Liquids", *Canadian Journal of Chemical Engineering*, Vol. 64, Iss. 5, pp. 839-843, 1986, https://doi.org/10.1002/cjce.5450640519

14. Ghanadzadeh, H., Ganji, M., Fallahi, S., "Mathematical Model of Liquid-Liquid Equilibrium for a Ternary System Using the GMDH-type Neural Network and Genetic Algorithm", *Applied Mathematical Modelling*, Vol. 36, Iss. 9, pp. 4096-4105, 2012, https://doi.org/10.1016/j.apm.2011.11.039

15. Pandharipane, S. L., Anish, M. S., Ankit, S., Sagar, G., "Modelling Combined VLE of Ten Binary Mixtures Using Artificial Neural Networks", *Proceedings of the International Conference on Intuitive Systems & Solutions*, 2012

16. Lashkarbolooki, M., Hezave, A. Z., Bayat, M., "Thermal Diffusivity of Hydrocarbons and Aromatics: Artificial Neural Network Predicting Model", Journal of Thermophysics and Heat Transfer, Vol. 31, No. 3, pp. 621-627, 2017, https://doi.org/10.2514/1.T5041

17. Pulga, L., Bianchi, G. M., Ricci, M., Cazzoli, G. et al., "Development of a Novel Machine Learning Methodology for the Generation of a Gasoline Surrogate Laminar Flame Speed Database under Water Injection Engine Conditions", *SAE International Journal of Fuels and Lubricants*, 2020, https://doi.org/10.4271/04-13-01-0001

18. Pulga, L., Bianchi, G. M., Falfari, S., Forte, C., "A Machine Learning Methodology for Improving the Accuracy of Laminar Flame Simulations with Reduced Chemical Kinetics Mechanisms", *Combustion and Flame*, 2020, https://doi.org/10.1016/j.combustflame.2020.02.021

19. Mohadesi, M., Moradi, G., Mousavi, H. S., "Estimation of Binary Infinite Dilute Diffusion

Coefficient Using Artificial Neural Network", *Journal of Chemical and Petroleum Engineering*, Vol. 48, No. 1, pp. 27-45, 2014

20. Liang, Z., Chen, L., et al., "Comprehensive chemical characterization of lubricant oils used in modern vehicular engines utilizing GC x GC-TOFMS", *Fuel*, Vol. 220, pp. 792-799, 2018, https://doi.org/10.1016/j.fuel.2017.11.142

21. Wen, Y., Bryan J., Kantzas, A., "Estimation of Diffusion Coefficients in Bitumen Solvent Mixtures as Derived From Low Field NMR Spectra", *Journal of Canadian Petroleum Technology*, Vol. 44, Iss. 4, 2005, https://doi.org/10.2118/05-04-03

22. Saltzman, E. S., King, D. B., Holmen, K., Leck, C., "Experimental Determination of the Diffusion Coefficient of Dimethylsulfide in Water", *Journal of Geophysical Research*, Vol. 98, Iss. C9, pp. 16481-16486, 1993, https://doi.org/10.1029/93JC01858

23. Sun, L., Meng, W., Pu, X., "New Method to Measure Liquid Diffusivity by Analyzing an Instantaneous Diffusion Image", *Optics Express*, Vol. 23, Iss. 18, pp. 23155-23166, 2015, https://doi.org/10.1364/OE.23.023155

24. Schatzberg, P., "Diffusion of Water Through Hydrocarbon Liquids", *Journal of Polymer Science*, Vol. 10, Iss. 1, pp. 87-92, 1965, https://doi.org/10.1002/polc.5070100108

25. Olson, R. L., Walton, J. S., "Diffusion Coefficients of Organic Liquids in Solution from Surface Tension Measurements", *Industrial & Engineering Chemistry*, Vol. 43, Iss. 3, pp. 703-706, 1951, https://doi.org/10.1021/ie50495a037

26. Reddy, K. A., Doraiswamy, L. K., "Estimating Liquid Diffusivity", *Industrial & Engineering Chemistry Fundamentals*, Vol. 6, No. 1, pp. 77-79, 1967, https://doi.org/10.1021/i160021a012

27. Moore, J. W., Wellek, R. M., "Diffusion Coefficients of N-Heptane and N-Decane in N-Alkanes and N-Alcohols at Several Temperatures", *Journal of Chemical and Engineering Data*, Vol. 19, No. 2, pp. 136-140, 1974, https://doi.org/10.1021/je60061a023

28. Vignes, A., "Diffusion in Binary Solutions. Variation of Diffusion Coefficient with Composition", *Industrial & Engineering Chemistry Fundamentals,* Vol. 5, No. 2, pp. 189-199, 1966, https://doi.org/10.1021/i160018a007

29. Fan, Y., Qian, R., Shi, M., Shi, J., "Infinite Dilution Diffusion Coefficients of Several Aromatic Hydrocarbons in Octane and 2,2,4-Trimethylpentane", *Journal of Chemical and Engineering Data*, Vol. 40, No. 5, pp. 1053-1055, 1995, https://doi.org/10.1021/je00021a004

30. Rodwin, L., Harpst, J. A., Lyons, P. A., "Diffusion in the System Cyclohexane-Benzene", *Journal of Physical Chemistry*, Vol. 69, No. 8, pp. 2783-2785, 1949, https://doi.org/10.1021/j100892a503

31. Sanni, S. A., Hutchinson, P., "Diffusivities and Densities for Binary Liquid Mixtures", *Journal of Chemical and Engineering Data,* Vol. 18, No. 3, pp. 317-322, 1973, https://doi.org/10.1021/je60058a028

32. Matthews, M. A., Akgerman, A., "Diffusion Coefficients for Binary Alkane Mixtures to 573 K and 3.5 MPa", *AIChe Journal*, Vol. 33, No. 6, pp. 881-885, 1987, https://doi.org/10.1002/aic.690330602

33. Funazukuri, T., Nishimoto, N., Wakao, N., "Binary Diffusion Coefficients of Organic Compounds in hexane Dodecane, and Cyclohexane at 303.2-333.2 K and 16.0 MPa", *Journal of Chemical and Engineering Data,* Vol. 39, No. 4, pp. 911-915, 1994, https://doi.org/10.1021/je00016a062

34. Lo, H. Y., "Diffusion Coefficients in Binary Liquids N-Alkane Systems", *Journal of Chemical and Engineering Data,* Vol. 19, No. 3, pp. 236-241, 1974, https://doi.org/10.1021/je60062a014

35. Shieh, J. J., Lyons, P. A., "Transport Properties of Liquid N-Alkanes", *Journal of Physical Chemistry*, Vol. 73, No. 10, pp. 3258-3264, 1969, https://doi.org/10.1021/j100844a017

36. Dymond, J. H., "Limiting Diffusion in Binary Nonelectrolyte Mixtures", *Journal of Physical Chemistry*, Vol. 85, No. 22, pp. 3291-3294, 1981, https://doi.org/10.1021/j150622a016

37. Kett, T. K., Anderson, D. K., "Ternary Isothermal Diffusion and the Validity of the Onsager Reciprocal Relations in Nonassociating Systems", *Journal of Physical Chemistry*, Vol. 73, No. 5, pp. 1268-1274, 1969, https://doi.org/10.1021/j100725a016

38. Perry, R. H., Green, D. W., "Perry's Chemical Engineers' Handbook 8th Edition", McGraw Hill, 2008

39. Cullinan, H. T., Toor, H. L., "Diffusion in the Three-Component Liquid System Acetone-Benzene-Carbon Tetrachloride", *Journal of Physical Chemistry*, Vol. 69, No. 11, pp. 3941-3949, 1965, https://doi.org/10.1021/j100895a050

40. Anderson, D. K., Hall, J. R., Babb, A. L., "Mutual Diffusion in Non-Ideal Binary Liquid Mixtures", *Journal of Physical Chemistry*, Vol. 62, No. 4, pp. 404-408, 1958, https://doi.org/10.1021/j150562a006

41. McKeigue, K., Gulari, E., "Effect of Molecular Association on Diffusion in Binary Liquid Mixtures", *AIChe Journal*, Vol. 35, Iss. 2, pp. 300-310, 1989, https://doi.org/10.1002/aic.690350215

42. Shuck, F. O., Toor, H. L., "Diffusion in the Three Component Liquid System Methyl Alcohol-N-Propyl Alcohol-Isobutyl Alcohol", *Journal of Physical Chemistry*, Vol. 67, No. 3, pp. 540-545, 1963, https://doi.org/10.1021/j100797a002

43. Chang, P., Wilke, C. R., "Some Measurements of Diffusion in Liquids", *Journal of Physical Chemistry*, Vol. 59, No. 7, pp. 592-596, 1955, https://doi.org/10.1021/j150529a005

44. te Riele, M. J. M., Snijder, E. D., van Swaaij, W. P. M., “Diffusion Coefficient at Infinite Dilution in Water and in N-Methylpyrrolidone”, *Journal of Chemical and Engineering Data*, Vol. 40, No. 1, pp. 34-36, 1995, https://doi.org/10.1021/je00017a009

45. Vadovic, C. J., Colver, C. P., “Infinite Dilution Diffusion Coefficients in Liquids”, *AIChE Journal*, Vol. 19, Iss. 3, pp. 546-551, 1973, https://doi.org/10.1002/aic.690190320

46. Baldauf, W., Knapp, H., “Experimental Determination of Diffusion Coefficients, Viscosities, Densities and Refractive Indexes of 11 Binary Liquid Systems”, Ber. Bunsenges. Phys. Chem., Vol. 87, Iss. 4, pp. 304-309, 1983, https://doi.org/10.1002/bbpc.19830870407

47. Poling, B. E., Prausnitz, J. M., O’Connell, J. P., “The Properties of Gases and Liquids 5th Edition”, McGraw-Hill 2000

48. King, C. J., Hsueh, L., Mao, K., “Liquid Phase Diffusion of Non-Electrolytes at High Dilution”, *Journal of Chemical and Engineering Data,* Vol. 10, No. 4, pp. 348-350, 1965, https://doi.org/10.1021/je60027a014

49. Yaws, C. L., “Thermophysical Properties of Chemicals and Hydrocarbons 2th Edition”, Gulf Professional Publishing, 2014

50. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., “Scikit-Learn: Machine Learning in Python”, *Journal of Machine Learning Research*, Vol. 12, pp. 2825-2830, 2011

51. Chollet, F., “Keras”, 2015, https://keras.io

52. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al., “Tensorflow: A System for Large-Scale Machine Learning”, 12th USENIX Symposium on Operating Systems Design and Implementation, pp. 265-283, 2016

53. Geron, A., “Hands-On Machine Learning with Scikit-Learn and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems”, O’Reilly, 2017

54. Hastie, T., Tibshirani, R., Friedman, J., “The Elements of Statistical Learning: Data Mining, Inference and Prediction”, Springer, 2009

55. Zhu, J., Zou, H., Rosset, S., Hastie, T., “Multi-Class AdaBoost”, *Statistics and its Interface*, Vol. 2, pp. 349-360, 2009, doi:10.4310/SII.2009.v2.n3.a8

56. Freund, Y., Schapire, R., “Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting”, *Journal of Computers and System Sciences*, Vol. 55, Iss. 1, pp. 119-139, 1997, https://doi.org/10.1006/jcss.1997.1504

57. Kingma, D. P., Ba, J. L., “Adam: A Method for Stochastic Optimization”, *Proceedings of the International Conference on Learning Representations*, 2014

58. Schulz, F., Beyrau, F., “Systematic Investigation of Fuel Film Evaporation”, *SAE Technical Paper* 2018-01-0310, 2018, https://doi.org/10.4271/2018-01-0310

59. Westbrook, C. K., Sjoberg, M., Cernansky, N. P., “A New Chemical Kinetic Method of Determining RON and MON Values for Single Component and Multicomponent Mixtures of Engine Fuels”, *Combustion and Flame*, Vol. 195, pp. 50-62, 2018, https://doi.org/10.1016/j.combustflame.2018.03.038

60. Malaguti, S., Bagli, G., Montanaro, A., Piccinini, S., “Experimental and Numerical Characterization of Gasoline-Ethanol Blends from GDI Multi-Hole Injector by Means of Multi-Component Approach”, *SAE Technical Paper* 2013-24-0002, 2013, https://doi.org/10.4271/2013-24-0002

61. Grunberg, L., Nissan, A. H., “Mixture Law for Viscosity”, *Nature*, Vol. 164, pp. 799-800, 1949, https://doi.org/10.1038/164799b0

62. Hiss, T. G., Cussler, E. L., “Diffusion in high viscosity liquids”, *AIChE Journal*, Vol. 19, Iss. 4, pp. 698-703, https://doi.org/10.1002/aic.690190404

Nomenclature

Abbreviations

DI: Direct Injection

DPI: Direct Port Injection

FFNN: Feed Forward Neural Network

HC: Hydrocarbon

LSPI: Low-Speed Pre-Ignition

ML: Machine Learning

NBP: Normal Boiling Point

NN: Neural Network

PFI: Port Fuel Injection

PI: Pre-Ignition

PM: Particulate Mass

PN: Particulate Number

RDE: Real Driving Emission

SI: International System of units

SPCCI: Spark Controlled Compression Ignition

TDC: Top Dead Center

TGDI: Turbo Gasoline Direct Injection

WLTC: Worldwide harmonized Light vehicles Test Cycle

Variables

A: association factor

ARE: Absolute Relative Error

C: carbons number

D: diffusion coefficient

H: hydrogens number

I: mutual information

K: generic constant

Lv: latent heat of vaporization

M: molecular weight

MARE: Mean Absolute Relative Error

O: oxygens number

p: generic probability

R^2: coefficient of determination

RMSE: Root Mean Squared Error

T: temperature

V: molar volume

V_b: molar volume at the normal boiling point

w: mass fraction

x: mole fraction

Greek letters

θ: machine learning prediction task

μ: dynamic viscosity

ρ: density

Φ: generic property

Subscripts

1: solvent

2: solute

mix: mixture

Appendix

Table A1. Rank of the input features for the Machine Learning input based on the mutual information analysis

Feature	Mutual Information
ρ_1	0.4261
V_1	0.3764
$\# H_1$	0.2935
$M_2^{0.5}$	0.2649
μ_1	0.2641
ρ_1/ρ_2	0.2422
ρ_2	0.2398
$\# C_1$	0.2350
M_1	0.2241
V_1/V_2	0.2202
Lv_1	0.2058
M_1/M_2	0.1950
M_2^2	0.1789
Lv_2^2	0.1663
V_2^2	0.1634
$\# C_2$	0.1532
ρ_2^2	0.1508
$\# O_1$	0.1439
$\rho_2^{0.5}$	0.1425
$\# O_2$	0.1418
$V_2^{0.5}$	0.1414
M_2	0.1401
V_2	0.1157
$Lv_2^{0.5}$	0.0965
$\# H_2$	0.0856

Table A2. Mass fractions (%) and properties for the tested gasoline surrogates and lubricant oils.

	I-octane	RON95	RD587	Ford-Synfuel	MIT	SAE10W30	SAE30
Cyclopentane	0	0	6.3736	0	0	0	0
i-pentane	0	0	0	16.61	20	0	0
n-Pentane	0	0	8.4304	0	0	0	0
Cyclohexane	0	0	0	24.05	0	0	0
1-Hexene	0	0	6.5558	0	0	0	0
n-Hexane	0	33.534	0	0	0	0	0
3-Methylpentane	0	0	0	0	10	0	0
Toluene	0	0	23.9247	17.65	13	0	0
n-Heptane	0	0	11.7079	0	5	0	0
m,p-Xylene	0	0	0	0	17	0	0
Ethylbenzene	0	0	0	12.93	0	0	0
i-Octane	100	45.24	43.0076	19.2	15	0	0
1,2,4-Trimethylbenzene	0	0	0	0	5	0	0
Naphthalene	0	0	0	1.09	0	0	0
n-Decane	0	21.226	0	8.47	3.5	0	0
i-Dodecane	0	0	0	0	1.5	0	0
Ethanol	0	0	0	0	10	0	0
Pentadecane	0	0	0	0	0	0.3	0
Hexadecane	0	0	0	0	0	0.5	0
Heptadecane	0	0	0	0	0	0.7	0
Octadecane	0	0	0	0	0	1	0
Nonadecane	0	0	0	0	0	1.5	0
Eicosane	0	0	0	0	0	2	0.2
Heneicosane	0	0	0	0	0	3	0.3
Docosane	0	0	0	0	0	4.5	0.4
Tricosane	0	0	0	0	0	5.5	0.7
Tetracosane	0	0	0	0	0	6.5	1
Pentacosane	0	0	0	0	0	7.5	1.3
Hexacosane	0	0	0	0	0	8.5	1.5
Heptacosane	0	0	0	0	0	9	1.9
Octacosane	0	0	0	0	0	8.5	2.6
Nonacosane	0	0	0	0	0	8	3.5
Triacontane	0	0	0	0	0	7.5	3.7
Hentriacontane	0	0	0	0	0	6.5	4.3
Dotriacontane	0	0	0	0	0	5.5	5.1
Tritriacontane	0	0	0	0	0	4	5.6
Tetratriacontane	0	0	0	0	0	3	6.2
Pentatriacontane	0	0	0	0	0	3	6.5
Hexatriacontane	0	0	0	0	0	1.5	6.7
Heptatriacontane	0	0	0	0	0	1	6.6
Octatriacontane	0	0	0	0	0	0.5	6.4
Nonatriacontane	0	0	0	0	0	0.3	5.9
Tetracontane	0	0	0	0	0	0.2	5.3
1-Hentetracontene	0	0	0	0	0	0	4.8
1-Dodetracontanethiol	0	0	0	0	0	0	4
Tritetracontane	0	0	0	0	0	0	3.4
2-Methyl-Tritetracontane	0	0	0	0	0	0	2.9
2-Methyltetra-Tetracontane	0	0	0	0	0	0	2.4
n-Heptatetracontane	0	0	0	0	0	0	1.9
n-Octatetracontane	0	0	0	0	0	0	0.9
n-Nonatetracontane	0	0	0	0	0	0	0.7
n-Pentacontan	0	0	0	0	0	0	0.5
n-Henpentacontane	0	0	0	0	0	0	0.3
n-Dopentacontane	0	0	0	0	0	0	0.2
Tripentacontane	0	0	0	0	0	0	0.1
Reference	/	[58]	[59]	[60]	[11]	[11]	[11]
M (kg/kmol)	114.23	110.77	98.97	97.59	93.45	387.48	497.97
ρ (kg/m^3)	692	689.28	729.88	761.25	746.83	875	890
μ (mPa· s)	0.503	0.487	0.452	0.583	0.509	113.75	311.50
L_v (kJ/kg)	272	323.98	336.23	363.44	411.21	257.38	221.23