Classification of fuel type for predictive maintenance in marine and industrial engines using time series feature extraction based on hypothesis

Applications in Energy and Combustion Science 25 (2026) 100440

Contents lists available at ScienceDirect

Applications in Energy and Combustion Science

journal homepage: www.sciencedirect.com/journal/applications-in-energy-and-combustion-science

Classification of fuel type for predictive maintenance in marine and industrial engines using time series feature extraction based on hypothesis tests and automated machine learning

Ning Guo [a,b,**], Erik Jansson [a], Mattias Johansson [a], Ronny Lindgren [a], Andreas Nyman [a], Jonas Sjöblom [c,*]

[a] AB Volvo Penta, Gropegårdsgatan 11, 417 15 Göteborg, Sweden
[b] 2550 Engineering AB, Qamcom Group AB, Falkenbergsgatan 3, 41285 Göteborg, Sweden
[c] Department of Mechanics and Maritime Sciences, Chalmers University of Technology, Chalmersplatsen 1, 41258 Göteborg, Sweden

ARTICLE INFO

Keywords:
Internal combustion engine
Fuel quality
Machine learning
Predictive maintenance
Time series
Tsfresh

ABSTRACT

Predictive maintenance in internal combustion engines can be enhanced by accurately identifying the fuel type based on data collected from sensors or electronic control units (ECUs). This paper presents a study that aims to predict the fuel type (HVO100 or EN590) using machine learning techniques, specifically based only on the engine's rotational speed. The rotational speed data of a heavy-duty 6-cylinder diesel engine is measured and downsampled to frequencies of 100, 1000, and 10,000 Hz. To extract relevant features from the time series data, hundreds of features are extracted using hypothesis tests via the tsfresh library. Subsequently, selected features are trained using Databricks' automated machine learning (AutoML) platform. The study explores the relationships between the number of features, downsampling frequency, and the choice of machine learning models. The results indicate that, under the current configuration, the best test F1 score of 0.995 is achieved using logistic regression with 20 features and a downsampling frequency of 10,000 Hz. The analysis of SHAP values and p-values revealed that components of the Fourier transform and wavelet transform of the rotational speed play crucial roles in distinguishing between the fuel types. It is our hypothesis that the differences observed in the frequency domain are related to variations in fuel characteristics. Overall, this study presents a simple, interpretable, and computationally cost-efficient machine learning solution for predicting fuel type in industrial engines. The findings demonstrate the potential of applying this approach in real-world production environments.

1. Introduction

In recent years, the advancement of Industry 4.0 has profoundly transformed the landscape of industrial operations, integrating Internet of Things (IoT) technologies, sophisticated data analytics and machine learning to enhance efficiency and predictive capabilities [1]. This enhanced levels of automation and data-driven decision making in industries, fostering improvements in operation efficiency and predictive maintenance [1]. One critical aspect is the deployment of IoT-enabled sensors in industrial machinery, which continuously collect and transmit data to centralized systems for analysis and actionable insights [2]. In the maritime and industrial engine manufacturing domain, Volvo Penta stands at the forefront of leveraging these technological advancements. The modern engines are equipped with sophisticated sensors and electronic control units (ECUs) that monitor various operational parameters and relay this data back to various locations, such as remote cloud servers, local storage environment, etc. This data-driven approach facilitates continuous monitoring and maintenance, ensuring optimal engine performance and longevity [3].

One way to unleash the potential of these data in the era of digitalization is to apply machine learning to determine the type of fuel used by end-users, eliminating the burden of direct communication of end-users. This is especially relevant when more users aim to be more environmentally friendly and transit from traditional diesel to biofuels like hydrotreated vegetable oil (HVO) [4]. Usually, engine

* Corresponding author.
** Corresponding author at: AB Volvo Penta, Gropegårdsgatan 11, 417 15 Göteborg Sweden.
E-mail addresses: ning.guo@consultant.volvo.com (N. Guo), jonas.sjoblom@chalmers.se (J. Sjöblom).

https://doi.org/10.1016/j.jaecs.2025.100440
Received 7 May 2025; Received in revised form 15 October 2025; Accepted 3 December 2025
Available online 3 December 2025

Nomenclature

a	Scaling parameter of wavelet transform
b	Translation (or positioning) parameter of wavelet transform
F	The set of all features
$f(x)$	Prediction using model f and data x
f_k	Corresponding frequency of the k^{th} component in Fourier transform
f_p	Peak frequency of a wavelet
f_s	Sampling frequency of original signal
i	Section 3.4.1: feature notation (feature i)Otherwise: imaginary number
k, m, n	The k^{th}, m^{th} and n^{th} element of a sequence, respectively
L	Loss function
N	Length of a sequence
p	P-value
Q	Desired false discovery rate
S	Any subset of the set of features F that does not include feature i (i.e., $S \subseteq F$ excluding i)
t	Time
th	Threshold
W	Wavelet transform
X	Fourier transform
x	Time series or sequential data
π	Mathematical constant that is the ratio of a circle's circumference to its diameter
$\pi\ (S)$	Kernel SHAP weight for subset S
Φ	SHAP value
ψ	Mother wavelet
ψ^*	ψ^* is the complex conjugate of the mother wavelet $\psi(t)$

manufacturers such as Volvo Penta does not have control over the type of fuel utilized by the end users in the engine. However, there is a widespread business interest in knowing the fuel type for several compelling reasons. Firstly, accurate fuel classification can lead to improved emission assessments, enhancing compliance with environmental regulations [5]. Secondly, different fuels have different characteristics and impacts on engine components, influencing maintenance needs and costs [6]. When the fuel type is known, it becomes easier to understand the long-term effects on an engine and its parts, such as oil, filters, and other components. This knowledge allows the impact of different fuels to be studied, and maintenance or part replacement needs to be predicted more accurately. As a result, service dates for engine parts, especially in the aftertreatment system, can be estimated more precisely based on the fuel type used. Additionally, knowing the fuel consumption behaviors facilitates better fuel supply chain management by predicting demand and optimizing logistics [7].

There are numerous relevant works done on this front. On one hand, a lot of effort was dedicated to map the properties of different fuels. Yadav et al. [8] used a heavy-duty single cylinder engine and tested four different fuels. In all load points of rated power, best brake thermal efficiency and cruise point, renewable fuels have lower emissions (CO_2, HC, CO, FSN) when compared to diesel fuels. Han et al. [9] performed experiments in a constant volume combustion chamber and compared the ignition and combustion characteristic of hydrotreated pyrolysis oil (HPO) to hydrotreated vegetable oil (HVO) and fatty acid methyl ester (FAME), diesel, and marine gas oil. According to them, HPO exhibits a longer ignition time compared to diesel-like fuels, whereas HVO has the shortest ignition delay time [9]. In addition, the blend of HVO and 75 % volume of HPO demonstrates similar combustion characteristics to diesel fuel, suggesting the possibility of creating a fully CO_2-neutral biofuel that performs very similarly in a compression ignition engine [9]. Khuong et al. [10] characterized the droplet evaporation of HVO under 573 – 873 K and 0.1 – 2.0 MPa, and they found that HVO droplet lives shorter and evaporates faster than diesel and they advised to pay attention to critical point of temperature and pressure for effective selection of evaporation conditions.

On the other hand, there are also advancements in applying machine learning in fuels and engines. In 2021, Aliramezani et al. [11] published a comprehensive review on the application of machine learning in internal combustion engines and suggested future directions to tackle challenges in the field. Wojcieszyk et al. [12] has developed a modelling approach to predict fuel consumption based on fuel properties, they achieved high accuracy for various fuel blends such as HCO, FAME, EN590, etc. Canal et al. [13] used ECU data to predict real-time fuel consumption and driving profile, their developed algorithm has lower computational cost and could conduct real-time analysis via cloud computing. Patil et al. [14] compared two approaches for predicting in-cylinder pressure in marine engines using data-driven models, and they found that the second approach, which utilizes a Fourier series function and artificial neural network regression, provides higher accuracy with a root mean square error within $\pm$ 0.2 bar when trained with 20 samples. Castresana et al. [15] applied ANN (artificial neural network) model on marine diesel engine and simultaneously predicted 35 performance and emission parameters with high accuracy.

Despite the impressive progress these researchers have made in applying machine learning in marine engines, there are several distinct challenges when applying these technologies in an industrial production environment as opposed to a development setting in a laboratory.

- Firstly, differentiating between closely related substances using available data can be challenging. Biofuels like HVO are intentionally designed to closely resemble traditional fuels such as diesel [16]. This similarity offers a practical benefit, allowing biofuels to replace diesel without significant infrastructure modifications [17]. However, it also presents a significant drawback: their inherent similarity makes it difficult to distinguish between them [16]. Additionally, the sensors and Electronic Control Units (ECUs) extensively implemented by manufacturers like Volvo Penta may not capture data that effectively indicate the optimal solutions to specific problems. For instance, while the exhaust emissions of certain chemical species might be critical indicators of the type of fuel combusted, the widespread installation of sensors to monitor these specific metrics could be either impractical or prohibitively expensive.
- Secondly, the machine learning models commonly employed for time series classification, such as Long Short-Term Memory (LSTM) networks [18] and Shape Dynamic Time Warping (ShapeDTW) [19], are often computationally intensive [20] and/or sometimes challenging to parallelize [21], which complicates their deployment in real-time industrial settings.
- Thirdly, interpreting the results of these machine learning models and correlating them with their physical and chemical significance remains a complex task, posing a barrier to fully leveraging IoT capabilities in enhancing operational efficiency and predictive maintenance.

To address the challenges presented, our research employs a dual methodology that combines time series feature extraction guided by hypothesis testing [22], with the efficiency of automated machine learning (AutoML) approaches. The following cornerstones build the originality of the work:

- This approach leverages data that are readily available from sensors/ECUs (in this case, rotational speed).
- By implementing the methodologies proposed in this study, one could deploy simpler and less resource-intensive machine learning models.
- The models not only yield robust predictions but also provide interpretable outcomes.

This integration of accessible data and efficient computational techniques is crucial for enhancing the operational efficacy and maintenance protocols in industrial settings, aligned with the advancements of Industry 4.0. Furthermore, this study seeks to resolve critical research questions listed below:

- Whether the type of fuel can be deduced from the rotational speed data captured by engine-installed sensors?
- How does sampling frequency influence prediction accuracy?
- Which characteristics of the rotational speed data are crucial for accurate fuel type prediction?

These inquiries are vital for advancing predictive capabilities and optimizing engine performance in the context of the ongoing digital transformation in industry. It should be noted that this work is a first-step feasibility study. Optimizing model performance is *beyond* the scope of this work.

The paper is structured as follows. This section gives the readers an overview of the background and motivations. Section 2 describes the combustion experiments that collected the rotational speed data. Section 3 presents the theory and methodology of data preprocessing, feature extraction and automated machine learning (AutoML). The produced results are discussed in Section 4 and then summarized in Section 5. In the end, Section 6 gives recommendations for future works.

2. Combustion experiment

The rotation speed data was generated using a Volvo D13K540, EU6SCR heavy duty 6-cylinder diesel engine [23], which is mounted on a dynamometer at Chalmers University of Technology in Sweden (cf. Fig. 1). The engine was operated at one load point, B50, corresponding to 1500 rpm and 50 % load. The combustion conditions were varied according to a full factorial Design of Experiment (DoE) using 3 parameters at 3 levels, in addition, we also tested a few extra rail pressures (cf. Table 1). The cylinder pressure in Cylinder No 2, out of 6 cylinders, was recorded, together with the rotational speed, recorded by an AVL angle encoder at 0.1 CAD resolution at 200 kHz. Two different fuels were used. One fossil diesel fuel (reference fuel for EN590) and one renewable diesel fuel (HVO100).

This experiment campaign was designed to serve multiple purposes and the current work is only one of them. The operating conditions relevant to this work is listed below in Table 1. It should be noted that both HVO100 and EN590 were subjected to testing under most operating conditions—including variations in rail pressure, injection timing, and pilot injection. However, certain conditions were tested using only one of the two fuels.

Table 1
Operating conditions of the combustion experiments for HVO100 and EN590.

Parameter	Option
Rail pressure (P_{rail}) [bar] (the underline marks the 3 DoE points)	723, 823, 923, 1023, 1223, 1423, 1623, 1823
Start of injection (*SOI*) [CAD (ATDC)]	−3, −6, −9
Pilot [μg/stroke]	0, 5, 10
Load point	B50 (1500 rpm, 50 % load)

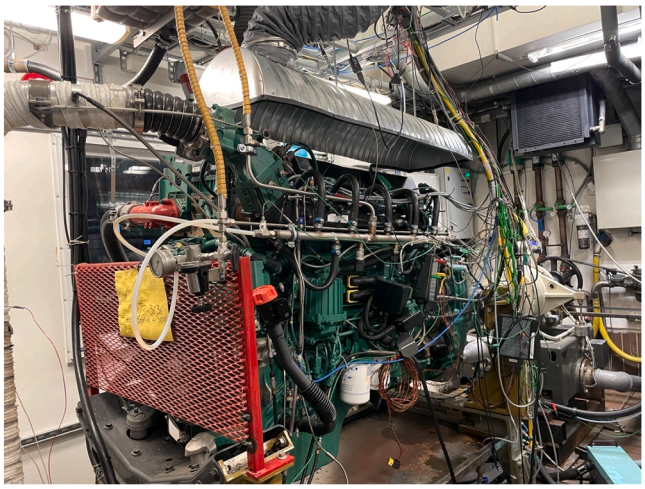

Fig. 1. Experimental setup featuring a Volvo D13K540 EU6SCR engine at Chalmers University of Technology. This figure provides a general overview of the rig configuration for illustrative purposes only and does not reflect the exact setup used in the experiments, such as wiring and other specific connections.

3. Theory and methodology

After obtaining measurements from combustion experiments in Section 2, the data was preprocessed and feature engineered before being sent to model training and result interpretation. This section explains these processes and Fig. 2 is the simplified flow chart for the readers to get an overview of these processes.

3.1. Data preprocessing

In the combustion experiment campaign (cf. Section 2), the sensors took a measurement every 5×10^{-6} s (i.e. 200,000 Hz) and produced relatively long time series data. Then the time series data was segmented into smaller chunks. The motivation for segmentation and chunking was to increase the number of samples for training, validation and testing. Short segments may not cover the entire engine cylinder combustion cycle and can be more susceptible to noise, while long segments might improve the model's recognition ability but also significantly increase computational costs, especially in resource-limited environments like ECUs or sensors in IoT systems. This is essentially a balancing act in real-world applications. Fig. 3 shows the rotational speed of four chunked samples from the experiment campaign to give the readers an intuitive view of the training data. The raw data collected by the sensor shows four peaks, and one peak indicates one rotation of 360°. It is consistent with the employed segmentation practice that each chunked sample consists of data for approximately two revolutions ($2 \times 720°$). The limited funding and time do not support long running experiments, so shorter segments were arbitrarily chosen to balance between having enough number of samples and keeping the integrity of combustion revolutions.

The raw data was downsampled to 10,000, 1000, and 100 Hz to as a pretext to explore how sampling frequency affects model performance. As shown in Fig. 3, the 10,000 Hz data still retained most characteristics of the raw data, including small variations and four distinct peaks. However, when the data was further downsampled to 1000 Hz, these four pronounced peaks were no longer visible, though some regular small variations remained. At 100 Hz, the data's variations were significantly smoothed out.

Based on these observations, it was decided to proceed only with the downsampled data for several reasons. First, the 10,000 Hz data already provide a sufficient representation of the raw data. Second, the raw data was sampled at a very high frequency, leading to large data size and exponentially increased computational demands, which are not feasible

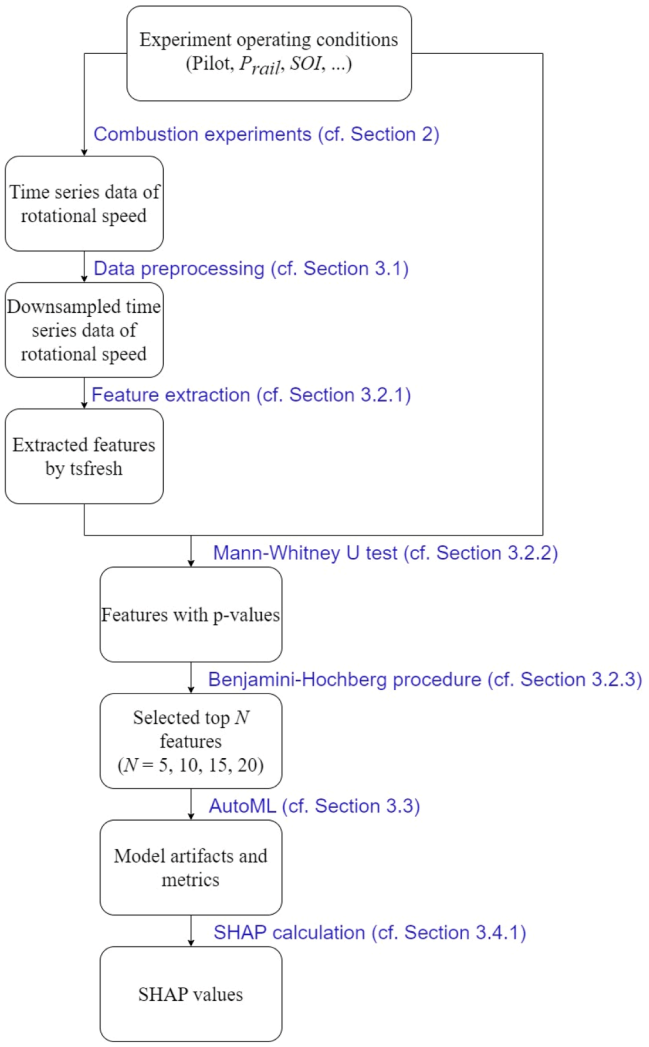

Fig. 2. Flow chart of the work process.

given the available resources and the limited business value. Third, most sensors/ECUs installed in commercially sold engines do not support such high resolutions. Therefore, even if results based on the raw data are promising, their applicability is limited, which contradicts our goal of conducting applied research for commercially feasible industrial applications where sampling frequency is low.

Essentially, it is a tradeoff between cost and accuracy when choosing sampling rate. Higher sampling rates can be used to improve model accuracy, but currently available rates are preferred for cost efficiency. In this study, a minimum sampling frequency of 100 Hz was selected, as it could be readily extracted from the existing engine control unit used in Volvo Penta's maritime engines. However, to aid the decision makers to determine if sensors of higher sampling rate are worth investing, 1000 Hz and 10,000 Hz were also examined in this study. By comparing results at different rates, the marginal benefits of increased sampling frequency are demonstrated, enabling stakeholders to decide on the most suitable approach in their business environment.

3.2. Feature engineering

Downsampled data from Section 3.1 was further processed through feature engineering before training through machine learning models. The methodology of feature engineering is closely aligned with the selection of machine learning models, especially in the context of time series classification, which can be tackled using a variety of approaches,

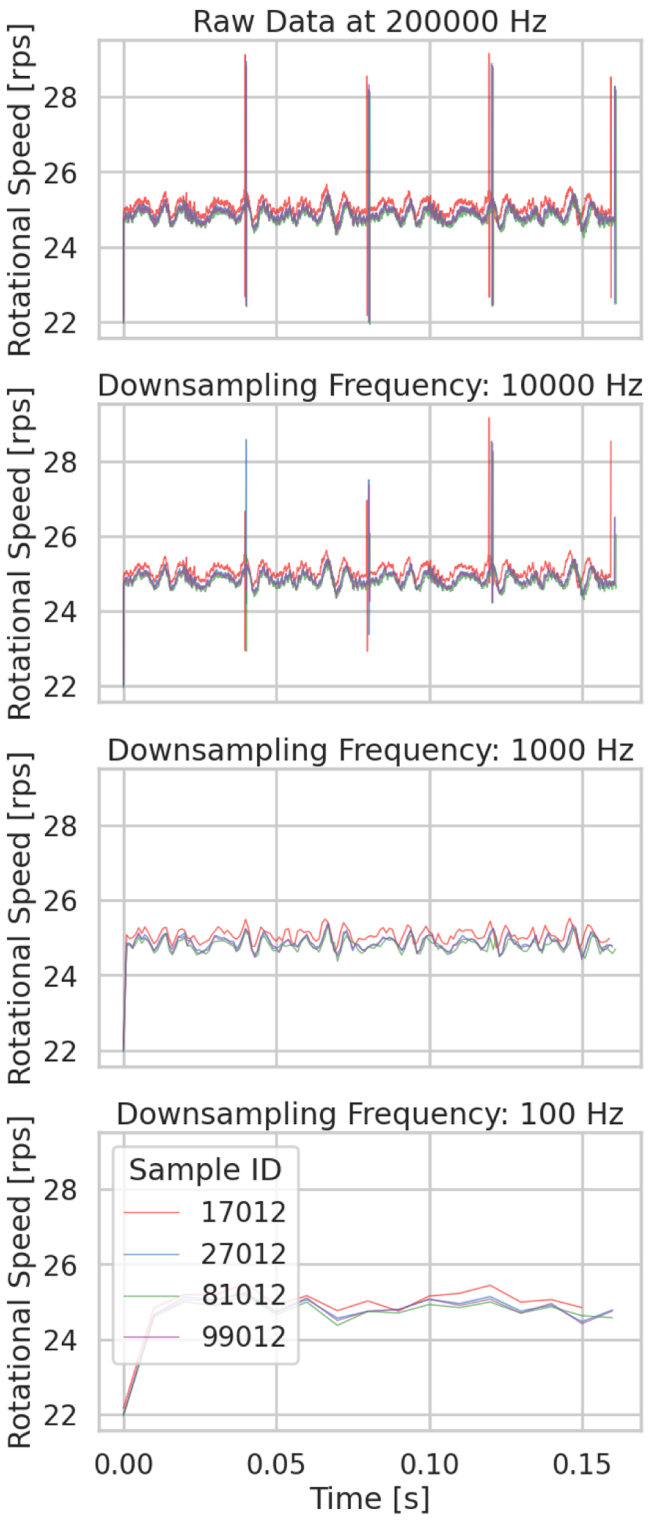

Fig. 3. Rotational speed [rps] of four samples. The first subplot is the raw data collected by the sensor, others are downsampled to 10,000, 1000 and 100 Hz. HVO100 was used for samples with ID 17,012 and 27,012, and EN590 was used for samples with ID 81,012 and 99,012. The sharp increase in the last subplot is due to scattered data points connected as a straight line in visualization.

each distinct in its application and benefits. The distance-based approach leverages distance metrics such as Dynamic Time Warping (DTW) to classify series by their similarity, proving intuitive and effective for datasets where the series' shape or similarity is crucial [19,24]. However, this method is computationally intensive and susceptible to

noise and timing shifts, rendering it less suitable for real-time applications [24,25]. Alternatively, the shapelet-based approach focuses on identifying small, representative sub-sequences within the data that are highly indicative of specific classes [24,26]. This method is valued for its high interpretability and effectiveness in recognizing local shape-based patterns, yet it demands considerable computational power and substantial training data [24]. Deep learning, employing neural networks like CNNs and RNNs, excels in extracting and learning the most predictive features automatically, adept at handling complex, non-linear interactions within the data [24,27]. Nonetheless, it requires extensive datasets and typically results in "black box" models, complicating result interpretation [24].

Based on the challenges and limitations outlined in Section 1, the feature-based methodology has been adopted. This approach is particularly favored as it yields interpretable results, operates efficiently with smaller datasets, and offers flexibility in feature selection [24,28]. It is particularly valuable in scenarios where limited computational resources and data availability are coupled with a need for high interoperability [24]. For this task, tsfresh (Time Series FeatuRe Extraction on the basis of Scalable Hypothesis tests), a robust and powerful open-source Python package, has been chosen from various options of feature-based methodology [22]. Through this package, the calculation of hundreds of time series characteristics and the evaluation of their relevance are enabled [22,29]. It has been utilized successfully by researchers across various disciplines, and reliable results have consistently been yielded [30–32]. The comprehensive theory and methodology of tsfresh are detailed in their original publications and source code [22,29]. However, the most crucial and relevant aspects are summarized and repeated here below for readers' convenience.

3.2.1. Feature extraction

Time series and other sequential data exhibit various characteristics such as the number of peaks, maximum values, autocorrelation, etc. These characteristics are referred to as features of the time series or sequential data in this context. The python package tsfresh automates the process of identifying and calculating these features, efficiently generating a comprehensive set of feature candidates for further processing [22,29]. To obtain an intuitive understanding, readers could refer to Table 2, where some selected features calculated by tsfresh are listed.

There are three predefined methods offered by tsfresh for extracting features from raw time series data: minimal, efficient, and comprehensive [22,29]. The minimal method includes a limited number of essential features for quick tests; the comprehensive method includes all available features for in-depth analysis; and the efficient method is a good option when runtime performance is important, as it provides most of the features from the comprehensive method but excludes computationally expensive features [22,29]. This work used the efficient method for feature extraction, as it can balance feature richness and computational efficiency and it is suitable for practical applications.

The extracted features from time-series measurement of rotation speed are counted in Fig. 4. The figure also includes four additional features to describe experiment operational conditions: rail pressure (P_{rail}), start of injection time (*SOI*), load point, and usage of pilot. As one can see, the downsampling frequency plays a crucial role in determining the number of features extracted from the time series measurement data of rotational speed. The general trend is that as the downsampling frequency decreases, there are also fewer features that could be extracted from the time series measurement data of rotational speed. This is because some of the feature calculators require a minimal length of the time series data, otherwise these feature calculators will return a NaN (Not a Number) value and the corresponding features will be discarded [22,29]. Additionally, the marginal effects of increasing the downsampling frequency become less pronounced. Increasing the downsampling frequency from 100 Hz to 1000 Hz leads to a substantial increase in the number of features extracted, from 383 to 700. However, further increasing the downsampling frequency from 1000 to 10,000 Hz results in a smaller gain of only 80 additional features, indicating diminishing marginal effects.

Table 2
A list of feature calculators and their corresponding brief descriptions that are relevant in the discussion (cf. Section 4). The complete list of features and detailed explanation could be found in original publications and source code of tsfresh [22,29].

Calculator Symbol [22,29]	Feature Description [22,29]
ar_coefficient	Unconditional maximum likelihood of an autoregressive process of the maximum lag
agg_linear_trend	Linear least-squares regression
augmented_dickey_fuller	Whether or not a unit root is present in a time series sample
autocorrelation	Autocorrelation of the given lag
change_quantiles	Choosing a specific range on a graph and finding the average amount of change between consecutive points within that range
cid_ce	CID: an efficient complexity-invariant distance for time series [7]
cwt_coefficients	Continuous wavelet transform based on Ricker wavelet (cf. Section 3.4.3)
energy_ratio_by_chunks	Ratio of the sum of squares of a specified segment to the sum of squares of the entire time series, dividing the series into a given number of segments
fft_coefficient	Fourier coefficients of discrete Fourier transform for real input [33] (cf. Section 3.4.2)
fourier_entropy	Binned entropy of the power spectral density using the Welch method [34]
mean_second_derivative_central	Mean value of a central approximation of the second derivative
number_peaks	Number of peaks at least support of specified value
partial_autocorrelation	Partial autocorrelation at the specified lag [35]
permutation_entropy	Permutation entropy [36]
spkt_welch_density	Cross power spectral density [34]

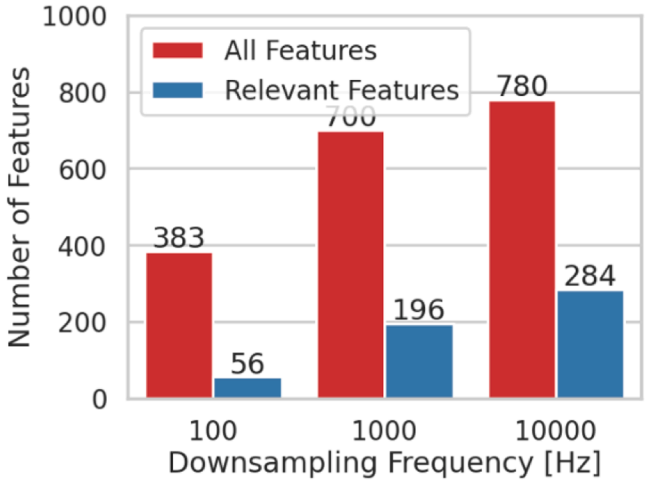

Fig. 4. Features from tsfresh extraction and experiment operating conditions across three downsampling frequencies. The relevance of features is determined by Mann-Whitney U test and Benjamini Hochberg procedure (discussed later in Section 3.2.3).

3.2.2. Mann-Whitney U test

All these features mentioned above (cf. Fig. 4), including rail pressure, *SOI* and the choice of pilot, were tested with Mann-Whitney U test [37,38]. In the context of machine learning, the Mann-Whitney U test can be utilized to assess whether the distributions of a feature differ significantly across two different groups (e.g., binary target variable). This test generates a p-value, which indicates whether the observed

differences in feature distributions between the groups are statistically significant. A low p-value suggests that it could be considered as a strong candidate for inclusion in the model for effective classification [22,29]. The nonparametric nature of the test makes it robust to non-normal data distributions, ideal for real-world datasets where assumptions of normality often do not hold.

Fig. 5 illustrates the p-values for all features, ranked in ascending order, to provide a clear understanding of feature relevance across different downsampling frequencies. Fig. 5 highlights that higher downsampling frequencies result in features with much lower p-values (the *y*-axis in Fig. 5 is in *logarithmic* scale), indicating much greater statistical significance. This is expected as samples with higher frequencies retain more signal detail. Additionally, the gradient for higher frequencies is much steeper. A steeper gradient suggests that significant features are identified more quickly as the number of features increases, enhancing differentiation between significant and non-significant features. Consequently, this improves the precision and reliability of predictive models, making higher downsampling frequencies more effective for predictive maintenance.

3.2.3. *Benjamini-Hochberg procedure*

After obtaining p-values for all these features, Benjamini-Hochberg procedure was performed to identify the relevant features [22,39]. The Benjamini-Hochberg procedure is a statistical method commonly used in hypothesis testing to mitigate the risks of the false discovery when conducting multiple statistical tests to determine the significance of individual features. This procedure is particularly relevant here as there are a large number of features (cf. Fig. 4), and conducting multiple tests simultaneously on large samples could lead to an increased chance of obtaining false positive results [22,39].

The Benjamini-Hochberg procedure addresses this issue by adjusting the p-values obtained from each individual test [22,39]. First, the p-values, $p_1, p_2, p_3, \ldots, p_m$ (m is the total number of hypotheses tested), are sorted in ascending order [22,39]. Then a significance threshold, or alpha level, is calculated as below:

$$p_{th,n} = \frac{n}{m} Q \tag{1}$$

Q is the desired false discovery (FDR) rate (0.05 by default in tsfresh), subscript n is the rank or the index of the sorted p-value and *th* means it is the threshold [22,39]. Then it identifies the largest p_n that could meet the criteria of $p_{th,n} < p_n$, assuming that it is p_k [22,39]. At last, it rejects the null hypotheses where the p-value is less than or equal to $p_{k\ 0}$ [22,39]. By controlling the FDR, the Benjamini-Hochberg procedure helps to identify the most relevant features while minimizing the number of false discoveries [22,39]. This is particularly important in feature selection for machine learning, as it ensures that the selected features are truly informative and contribute to the predictive performance of the model [22,39].

The number of features that are considered relevant by the Benjamini-Hochberg procedure is plotted previously in Fig. 4 and the corresponding proportion of relevant features out of all features is shown in Fig. 6. As the downsampling frequency increases, not only the number of features but also the proportion of relevant features increases. This suggests that higher downsampling frequencies provide more detailed and significant information from the data, consistent with the previous discussions. However, there is also evidence of diminishing marginal effects, as the rate of increase slows down as the downsampling frequency gets higher. The trend of diminishing marginal effects highlights the importance of finding an optimal balance between downsampling frequency and the proportion of relevant features. While higher frequencies capture more significant information, the incremental benefits decrease, suggesting a point of optimal frequency beyond which further increases yield low returns.

A heuristic decision is made to consider only the top 20 relevant features for each downsampling frequency, given that the steep gradients on the left side of Fig. 5 indicate that the top few features are much more relevant than most of the others. Their p-values are presented in Fig. 7. Across various downsampling frequencies, certain features consistently emerge as relevant, although their importance ranking may vary, for example, spkt_welch_density [34]. Each downsamplineg frequency also highlights unique features not relevant at other frequencies. At a downsampling frequency of 100 Hz, the features — selected based on p-values — examine the time series data for patterns and properties inherent to the time domain, such as trends, autocorrelation, stationarity, distributional shifts, etc. In contrast, at a downsampling frequency of 10,000 Hz, the relevant features are primarily concerned with transforming and analyzing the data to extract frequency-based information, moving away from time domain analysis or statistical attributes. Additionally, their p-values are notably lower. This is consistent with the discussion about the raw and downsampled data in Fig. 3.

It's important to recognize that certain features can be redundant. For example, the real and imaginary parts of the Fourier transform together provide the same information as the combination of the magnitude and angle derived from the Fourier transform results. We would like to clarify that the primary objective of this work is to develop

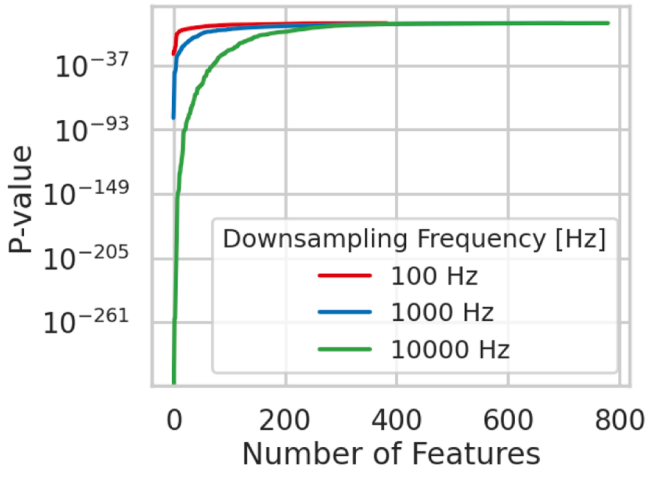

Fig. 5. P-value (in *logarithmic* scale), in ascending order, produced by the Mann-Whitney U test [37,38] for each feature across three downsampling frequencies.

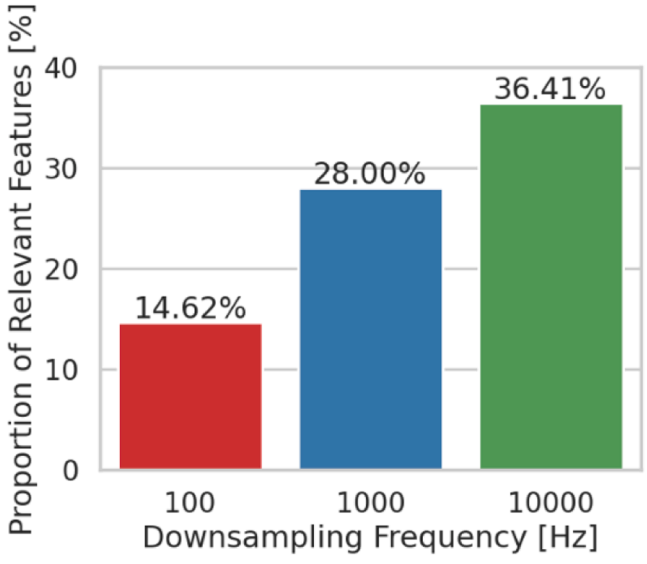

Fig. 6. Proportion of relevant features by downsampling frequency using the Benjamini-Hochberg procedure.

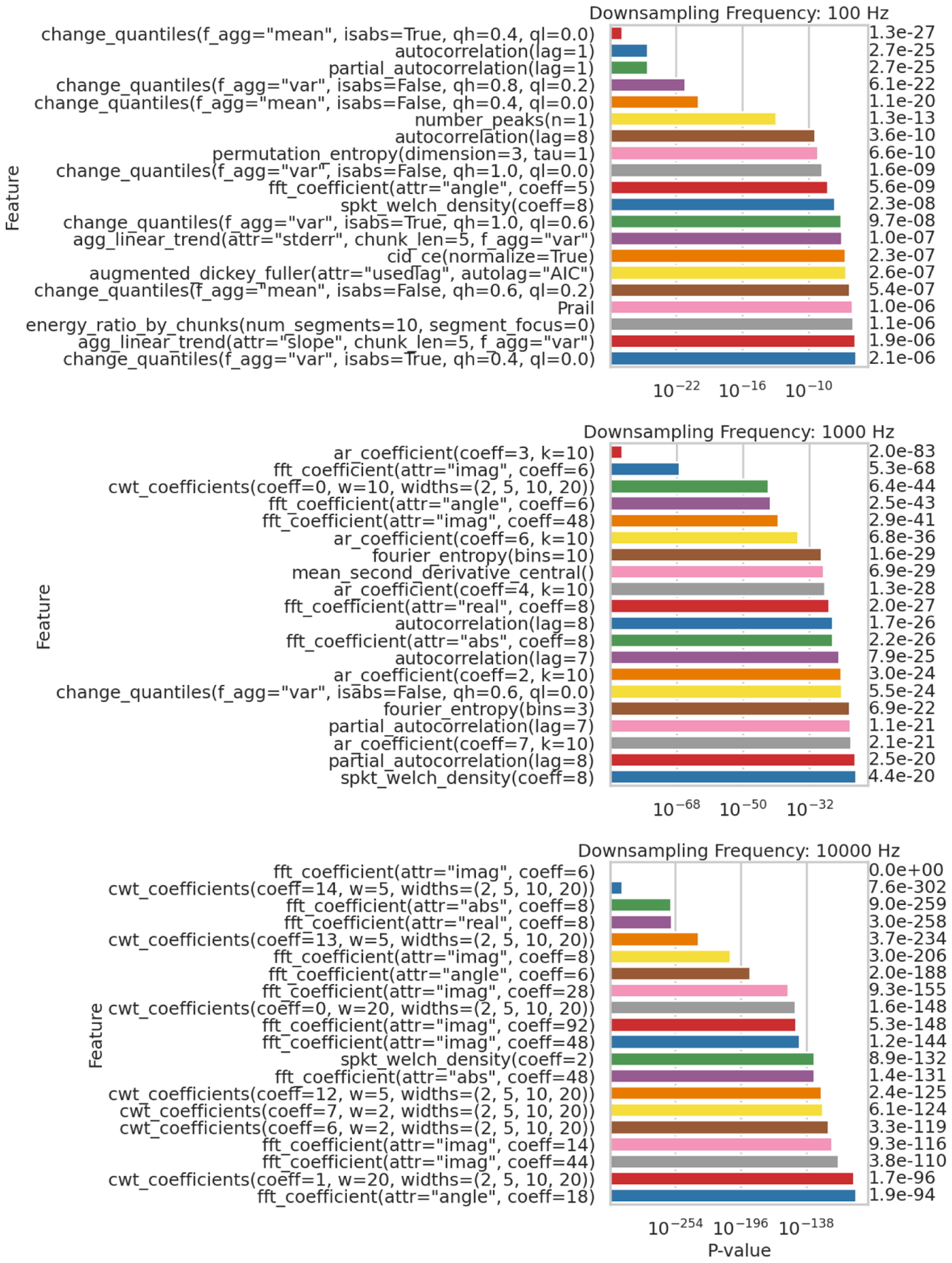

Fig. 7. Top 20 relevant features based on p-values across three sampling frequency. The description of features can be found in Table 2.

a rapid proof-of-concept to validate the feasibility of our proposed method, rather than to optimize model performance or computational efficiency at this stage. As such, our focus has been on establishing an automated pipeline that can robustly select relevant features with minimal manual intervention. Identifying and removing redundant features often requires domain expertise and manual input, which is the future works but beyond the scope of this initial proof-of-concept. Additionally, while certain feature combinations (e.g., $a + b$ and $c + d$) may be information equivalent, the individual features (a, b, c, or d) are not necessarily interchangeable or equivalent. Retaining these features allows for more granular analysis and interpretation in subsequent studies. Furthermore, many machine learning algorithms, such as those employing tree-based methods, are inherently capable of handling redundant features during model training. In the current experiments, the presence of redundant features could still achieve satisfactory results. Nevertheless, the development of automated methods for identifying and removing redundant features shall be a key focus of the future works to optimize model performance and computational efficiency.

3.3. Automated machine learning (AutoML)

The research began with data preprocessing, feature engineering and selection, preparing the data for the subsequent machine learning phase. Given the exploratory objectives outlined in Section 1, Automated Machine Learning (AutoML) was employed to streamline and expedite the process [40]. AutoML automates key tasks such as feature engineering, model selection, hyperparameter tuning, and model evaluation, which are integral to developing and optimizing machine learning models [40]. AutoML is particularly advantageous for Exploratory Data Analysis (EDA), a process that involves analyzing and visualizing data to uncover insights, patterns, and relationships. By automatically generating and assessing various machine learning models, AutoML facilitates the identification of significant features, selection of the most effective model architecture, and optimization of hyperparameters to maximize performance. This approach not only accelerates the exploratory process but also enhances the overall effectiveness of the analysis.

AutoML begins by cleaning and preparing the data for training, then explores a variety of machine learning algorithms to address different problem types [40]. It performs distributed model training and hyperparameter tuning across these algorithms to identify the most effective model for the given task [40]. Using common evaluation metrics such as test F1 score, AutoML evaluates the performance of each algorithm and selects the best-performing model, ultimately presenting the optimized model and its configuration to the user [40].

This study utilized Databricks' AutoML [41]. It was operated on the Databricks runtime 14.3 LTS ML, which includes Apache Spark 3.5.0, GPU support, and Scala 2.12. As shown in Fig. 8, the AutoML experiment campaign tested two variables: the downsampling frequency (100, 1000, and 10,000 Hz as in Fig. 3) and the number of features (top 5, 10, 15, and 20 features as identified from Fig. 7), leading to a total of 12 AutoML experiments. In each experiment, Databricks' AutoML evaluated five established machine learning models — Decision Tree [42], Random Forest [42], Logistic Regression [42], XGBoost [43], and LightGBM [44] — across multiple runs, each with varying hyperparameters, as shown in Fig. 8. The selection of models and hyperparameters for each run was guided by Databricks' proprietary algorithm. The samples were stratified and divided into 60 %, 20 %, and 20 % approximately for training, validation, and testing, respectively. There are 2155 samples used in training, 695 samples used in validating and 727 in testing. Each experiment was configured with a 15-minute time out.

3.4. Feature and model interpretation

After AutoML is applied, the interpretation of the models and the identification of important features are addressed. This process involves two main steps. First, relevant features are determined using SHAP value analysis, as described in Section 3.4.1, and p-value analysis. Next, these key features are correlated with domain knowledge to enhance understanding of their physical and chemical significance. Since features derived from both Fourier and wavelet transforms are identified as important, brief theoretical backgrounds for these techniques are

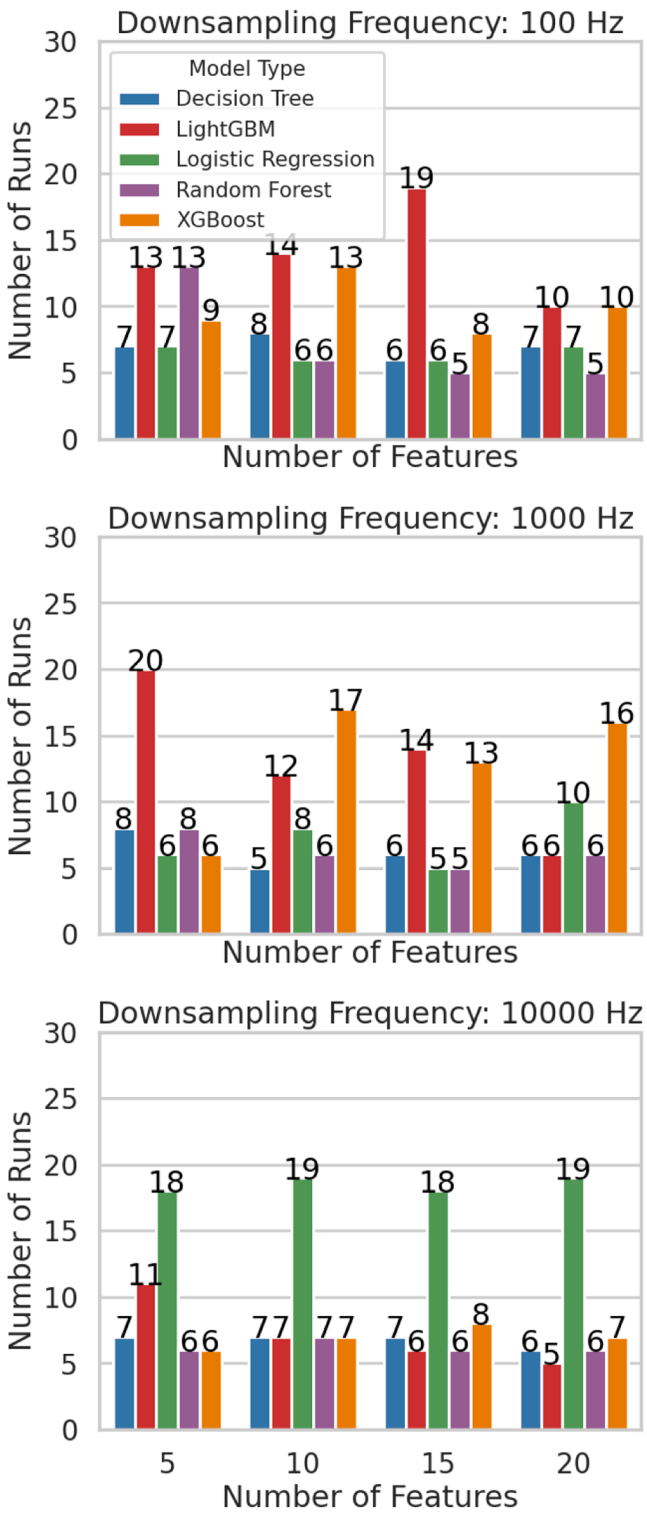

Fig. 8. Design of AutoML experiments. There are 12 experiments, each with its unique set of downsampling frequency (from top to bottom: 100, 1000 and 10,000 Hz) and number of features (from right to left: 5, 10, 15 and 20). A certain number of runs were tried for five different models in each experiment.

provided in Sections 3.4.2 and 3.4.3 for the readers' convenience.

3.4.1. SHAP values

SHAP (SHapley Additive exPlanations) values have emerged as a powerful tool for interpreting the predictions of machine learning models [45]. They provide a unified framework for explaining individual predictions by quantifying the contribution of each feature to the model's output [45]. SHAP values are based on cooperative game theory, specifically the Shapley value, which assigns a unique contribution to each feature based on its marginal impact on the prediction [45]. By decomposing the prediction into the contributions of individual features, SHAP values offer insights into the model's decision-making process and help identify which features are most influential [45]. It is calculated as follows:

$$\phi_i(f,x) = \sum_{S \subseteq f \text{ excluding } i} \frac{|S|!\cdot(|F|-|S|-1)!}{|F|!}[f(x_{S\cup i}) - f(x_S)] \tag{2}$$

where:

- $\phi_i(f, x)$ is the SHAP value for feature i;
- F is the set of all features;
- S is any subset of the set of features F that does not include feature i (i.e., $S \subseteq F$ excluding i);
- $|S|$ is the number of elements in the subset S;
- $f(x_S)$ is the model's prediction when only the features in subset S are known, with all other features being unknown or set to a baseline value;
- $f(x_{S\cup i})$ is the model's prediction when the features in subset S and feature i are known;
- $|F|!$ represents the factorial of the total number of features n; and
- $|S|!$ is the factorial of the size of subset S, and $(|F|-|S|-1)!$ is the factorial of the number of features not in S or i.

In the context of this research, kernel SHAP is used, as it is more computationally efficient while able to maintain similar level of accuracy [45]. Kernel SHAP assigns a weight, $\pi(S)$, to each subset S based on the number of features in the subset $|S|$:

$$\pi(S) = \frac{|F|-1}{|S|\cdot(|F|-|S|)} \tag{3}$$

Kernel SHAP solves a weighted linear regression problem to find the SHAP values ϕ_i by minimizing the following loss function:

$$L = \sum_{S\subseteq F} \pi(S)\cdot\left[f(x_s) - \left(\phi_0 + \sum_{i\in S}\phi_i\right)\right]^2 \tag{4}$$

where ϕ_0 is the base value without any features.

3.4.2. Fourier transform

The Fourier transform is a mathematical technique used to analyze signals and represent them in the frequency domain. It decomposes a signal into its constituent frequencies, revealing the amplitude and phase information of each frequency component. The Fourier transform operates on continuous or discrete signals and provides a global view of the frequency content of the entire signal. It uses a set of complex exponential functions as basis functions to represent the signal. The resulting spectrum obtained from the Fourier transform represents the signal's frequency content and can be used for various applications such as filtering, compression, and spectral analysis.

In the tsfresh algorithm, the fft_coefficient feature calculator uses discrete Fourier transform [33]:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i2\pi\frac{k}{N}n} \tag{5}$$

It transforms a sequence of $(x_n \mid n = 0, 1, 2, \ldots, N-1)$ in time domain to a sequence in frequency domain $(X_k \mid n = 0, 1, 2, \ldots)$. If f_s is the sampling frequency of original sequence, $(x_n \mid n = 0, 1, 2, \ldots, N-1)$, then the k^{th} component X_k corresponds to the following frequency:

$$f_k = \frac{k}{N} f_s \tag{6}$$

3.4.3. Wavelet transform

The wavelet transform is a mathematical tool used for signal analysis that provides a localized view of the frequency content of a signal [46]. Unlike the Fourier transform, which uses fixed basis functions, the wavelet transform uses wavelet functions that are localized in both time and frequency domains [46]. This allows the wavelet transform to capture both transient and stationary features of a signal effectively [46]. The wavelet transform decomposes a signal into a set of wavelet coefficients at different scales and positions, providing a time-frequency representation of the signal [46]. It is particularly useful for analyzing non-stationary signals with time-varying frequency content [46].

The Continuous Wavelet Transform (CWT) of a signal x (t) is defined as

$$W(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(t)\psi^*\left(\frac{t-b}{a}\right)\mathrm{d}t \tag{7}$$

Where a is the scaling parameter, b is the translation parameter, which controls the location of the wavelet. $W(a, b)$ is the wavelet coefficient at scale a and position b, and ψ^* is the complex conjugate of the mother wavelet $\psi(t)$. In the context of tsfresh, the wavelet transform is computed for a discrete set of scales a_n and positions b_n on the signal $x(t)$, which is sampled at discrete times t_n. The wavelet coefficients are computed as:

$$W(a_n,b_n) = \sum_{n=0}^{N-1} x(t_n)\psi^*\left(\frac{t_n-b_n}{a_n}\right)\Delta t_n \tag{8}$$

Here, the Ricker wavelet [47,48] is used:

$$\psi(t) = \frac{2}{\sqrt{3\sigma\pi^{1/4}}}\left(1-\frac{t^2}{a^2}\right)e^{\frac{-t^2}{2a^2}} \tag{9}$$

Where its peak frequency, f_p, is related to the scale, a, as

$$f_p = \frac{1}{\sqrt{2}\pi a\Delta t} \tag{10}$$

4. Result and discussion

4.1. Model performance from AutoML experiments

To evaluate the results from the AutoML experiments, three metrics commonly used in machine learning were plotted in Fig. 9: F1 score, accuracy, and ROC AUC (Receiver Operating Characteristic - Area Under the Curve) [49]. Accuracy measures the proportion of correctly classified instances out of the total, offering a quick overview of model performance, but it can be misleading in cases of class imbalance as it does not differentiate between types of errors [49]. ROC AUC is a robust metric for evaluating binary classifiers, measuring the model's ability to discriminate between positive and negative classes [49]. The F1 score is a reliable metric for binary classification performance as it considers both precision and recall, providing a balanced assessment that accounts for false positives and false negatives [49]. It provides a comprehensive evaluation across all possible classification thresholds, summarizing the trade-off between true positive and false positive rates [49]. These three metrics together provide a well-rounded assessment of model performance [49]. As one can infer from Fig. 9, the general trend is that more features and higher downsampling frequency leads to better scoring across these three metrics. Additionally, the marginal benefits of increasing downsampling frequency by a factor of 10 are higher than

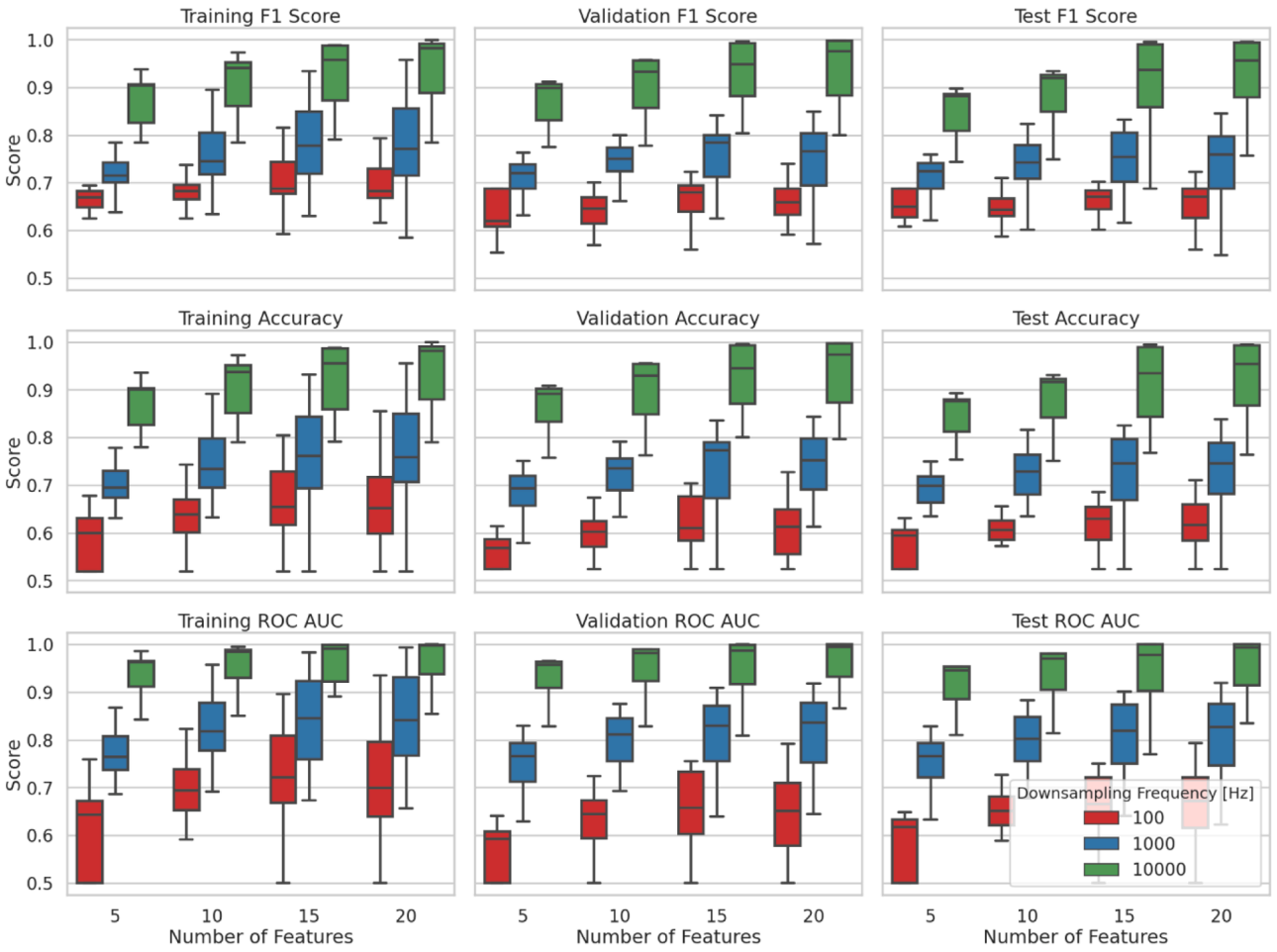

Fig. 9. F1, accuracy and ROC AUC scores for training, validation, and test datasets of all runs from the AutoML experiments for different downsampling frequency and number of features. The horizontal line denotes 0th, 25th, 50th, 75th and 100th percentiles, respectively, from the bottom to the top of the boxplot.

adding 5 more features. This will be discussed in detail in Fig. 11.

As previously mentioned, Databricks' AutoML trains and evaluates classification models using the following five algorithms: Decision Tree [42], Random Forest [42], Logistic Regression [42], XGBoost [43], and LightGBM [44]. The performance metrics assessed include F1 score, accuracy, and ROC AUC scores across various models, as illustrated in Fig. 10. The results show variability in model performance with no consistent leader or laggard across all experiments. Logistic Regression achieved the highest F1 scores in experiments with 15 and 20 features at 10,000 Hz, while Decision Tree generally performed poorly, especially in experiments of 1000 and 10,000 Hz with 20 features. LightGBM and XGBoost showed a range of outcomes, sometimes scoring both highest and lowest in the same experiments, which may account for the variability seen in their results. Notably, Random Forest did not record the highest F1 scores in any experiment but was the lowest at 10,000 Hz with 10, 15, and 20 features. This variability suggests that the choice of model and feature set should be tailored to specific use cases, as no single model consistently outperforms others across all settings. The significant variations in performance, particularly with LightGBM and XGBoost, underline the importance of detailed evaluation across different conditions to identify the most effective algorithm for a given scenario.

The runs with best F1 score for the test dataset for each experiment are shown in Fig. 11. The highest F1 scores are observed at the highest sampling frequency (10,000 Hz) across all feature sets, with the scores being very close to 1 for 15 and 20 features (around 0.995). This near-perfect score suggests very high accuracy under these conditions. The lowest F1 score occurs at the lowest sampling frequency (100 Hz) with the least number of features (5), which is still a moderate score of 0.688, indicating reasonable performance but room for improvement. It is consistent with the aforementioned general trend that increasing number of features and higher downsampling frequency led to higher best test F1 scores from AutoML. There is an exception at 100 Hz when number of features increases from 10 to 15. It appears that increasing the downsampling frequency tends to have a more substantial impact on improving the F1 scores compared to increasing the number of features, especially when the initial number of features is low. However, as the number of features becomes higher, the incremental benefits of adding more features might diminish. One should also consider the trade-off between increasing the number of features by 5 more and increasing the sampling frequency by 10 times. Increasing the sampling frequency by 10 times (e.g., from 100 Hz to 1000 Hz, or 1000 Hz to 10,000 Hz) significantly increases the amount of data collected and processed. However, the increase in data volume when adding 5 more features is generally less drastic. Additionally, high-frequency data collection might not always be scalable or practical, especially in resource-constrained environments or in applications where data collection is expensive or logistically challenging. Depending on the specific application, the sensitivity to changes in sampling frequency or number of features can vary. For instance, in real-time systems or battery-operated devices, the increased power consumption and processing needs for higher data sampling rates might not be feasible.

4.2. *Feature relevance: a retrospective*

The evaluation of feature relevance employs two distinct metrics: p-

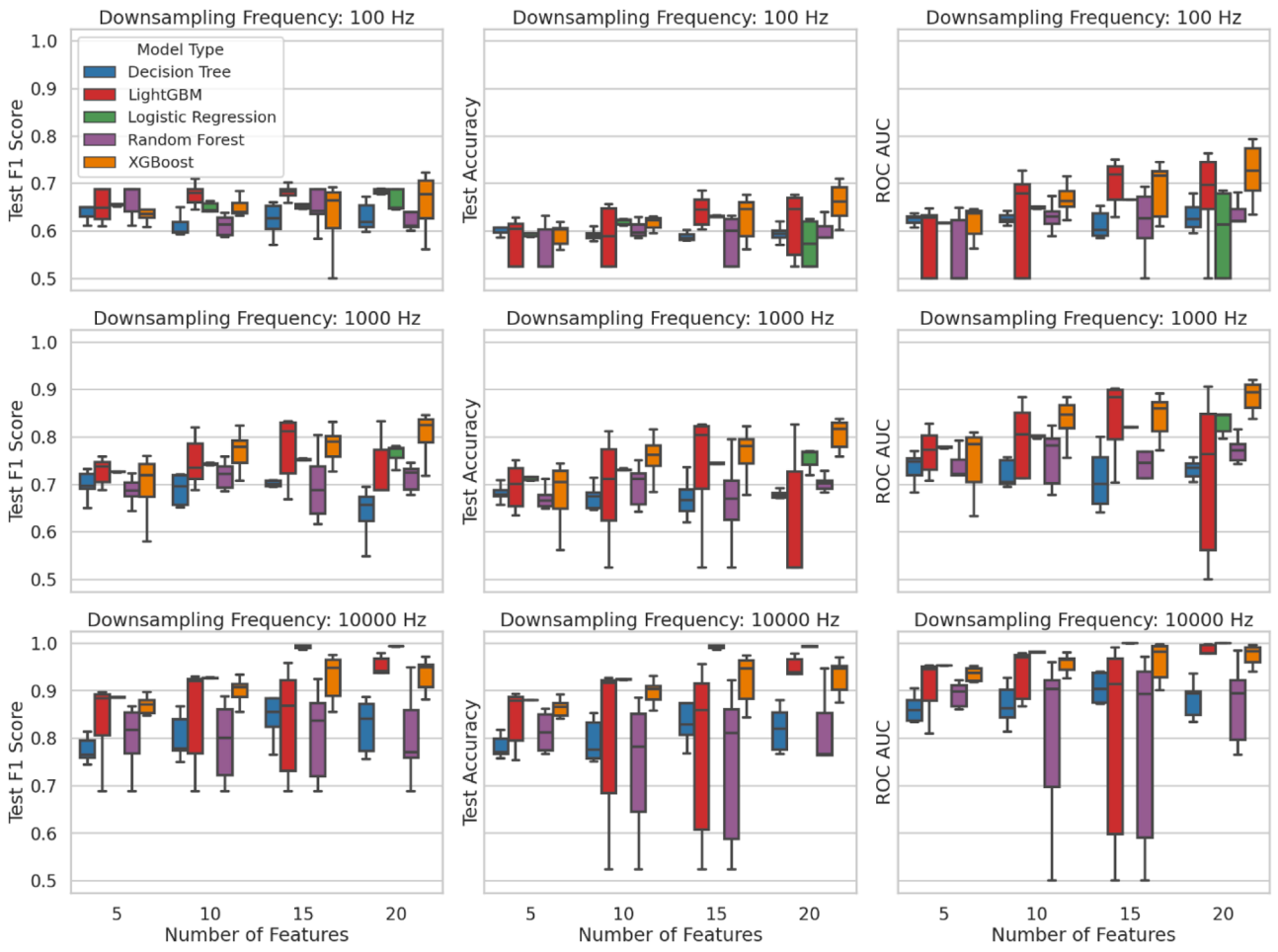

Fig. 10. F1 score, accuracy and ROC AUC for the test dataset for different models used in AutoML. The horizontal line denotes 0th, 25th, 50th, 75th and 100th percentiles, respectively, from the bottom to the top of the boxplot.

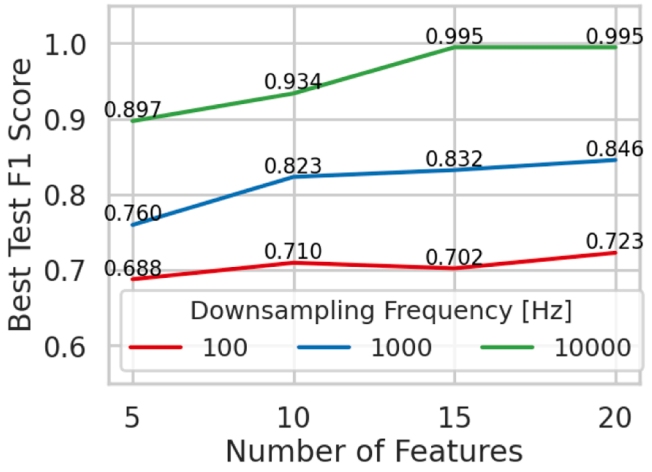

Fig. 11. The best F1 score for the test datasets for each experiment in AutoML.

values and SHAP values. P-values are determined prior to the phases of model training, validation, and testing to assess the statistical significance of features, while SHAP values are calculated post these phases to quantify the impact of each feature on model predictions. Fig. 12 illustrates the SHAP values for all 20 features across various models at a downsampling frequency of 10,000 Hz, with features organized in ascending order of their p-values as referenced in Fig. 7. This arrangement reveals differing levels of feature importance across models, as indicated by the variable lengths and distributions of the SHAP value dots. Such variations underscore the need for model-specific feature analysis, demonstrating that models interpret and weigh the same features differently. Notably, although features with the lowest p-values do not always correspond with the highest SHAP values, certain features consistently display high mean absolute SHAP values and low p-values across multiple models, suggesting their pivotal role in prediction across different modeling approaches. This is particularly evident with features like fft_coefficient(attr = "img", coeff = 6) and cwt_coefficient (coeff = 14, w = 5, width = (2, 5, 10, 20)).

Despite discrepancies in SHAP values between models, the influential direction of a given feature generally remains consistent, meaning a feature that positively affects the predictions in one model tends to do so across others. For instance, all five models demonstrate that lower values of fft_coefficient(attr = "img", coeff = 6) positively influence the predictions, indicating the classification label to be “True” (i.e. HVO100). Meanwhile, lower values fft_coefficient(attr = "img", coeff = 48) negatively affect the predictions. These features will be discussed in detail later in Section 4.3.

Each feature extracted from the Fourier and wavelet transforms, namely fft_coefficient and cwt_coefficients, converts data from the time domain to the frequency domain, enriching the representation of the dataset. Fig. 11 reveals that using just five features from these transforms at a 10,000 Hz frequency can achieve a 0.897 test F1 score, suggesting that even a limited frequency domain representation can yield robust prediction outcomes. Moreover, an increase in the number of features tends to enhance the test F1 score, further indicating that a

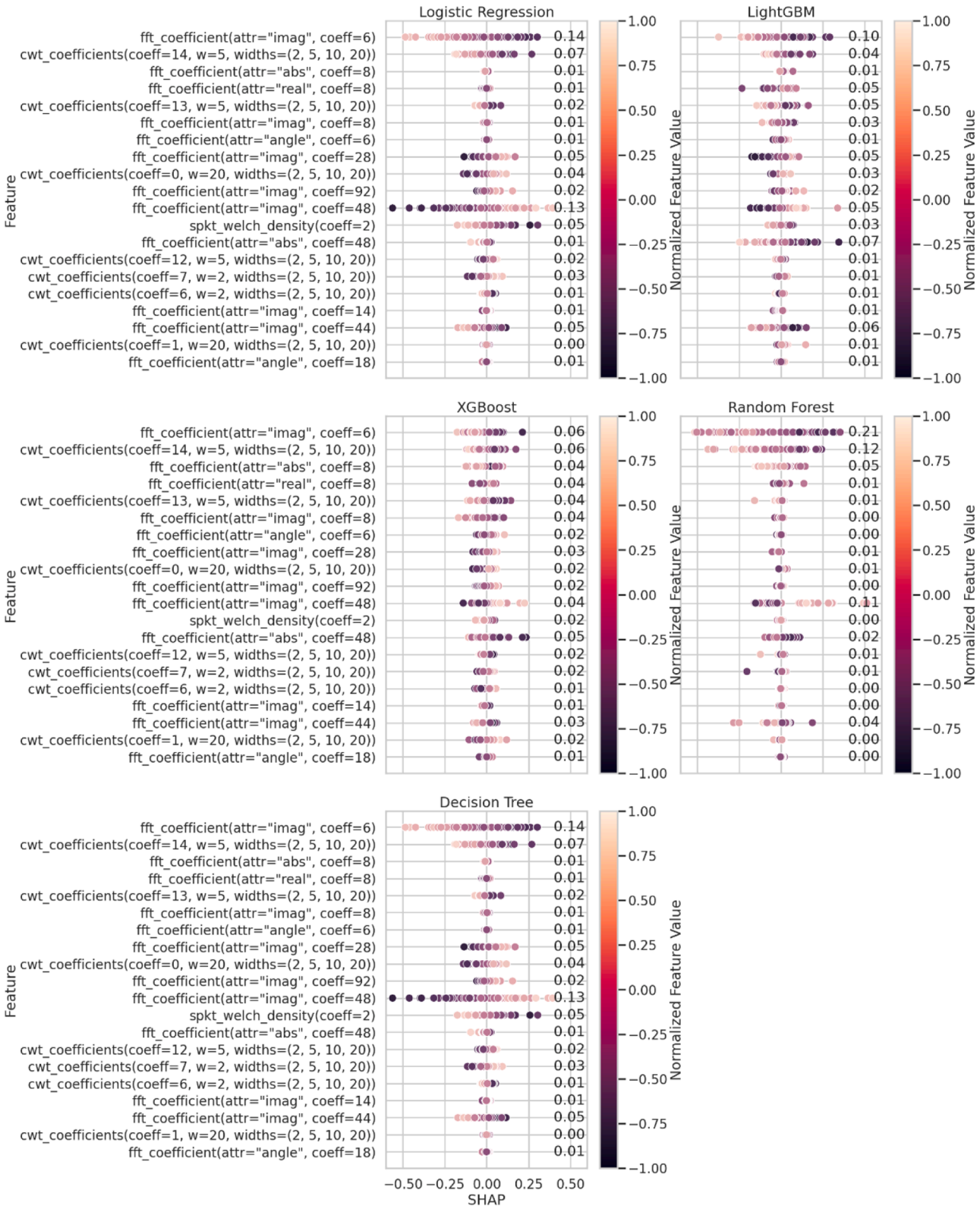

Fig. 12. Impact of top 20 features on model predictions across five machine learning models. This figure displays SHAP values for the top 20 features determined by the lowest p-values, at a downsampling frequency of 10,000 Hz. The *x*-axis represents SHAP values from a kernel explainer, which quantifies each feature's influence on the model's predictions. To the right of each subplot, the mean absolute SHAP value is shown, providing a measure of each feature's overall importance for each model. Features on the *y*-axis are sorted by increasing p-value from top to bottom, highlighting their significance according to Fig. 7. The SHAP value was generated by 500 samples (default configuration of Databricks).

more comprehensive representation in the frequency domain can lead to improved model performance. This insight highlights the effectiveness of frequency domain features in capturing essential information for accurate model predictions.

The analysis suggests that features derived from the frequency domain are highly effective for classification, as evidenced by their robust predictive performance. This is supported by the normalized Kernel Density Estimate (KDE) plots shown in Fig. 13, where the filled plots demonstrate how these features contribute to accurate predictions. The white line within each plot, delineating the blue and red areas, serves as an indicator of the ease with which two classes can be differentiated based on a specific feature value. The steepness of the white line is particularly telling; a steeper line indicates clearer class distinction. For instance, the white line for fft_coefficient(attr = "img", coeff = 6) is quite steep, suggesting that this feature is highly discriminative, which is corroborated by its very low p-value (cf. Fig. 7). Furthermore, the orientation of the plot aligns with the SHAP value results from Fig. 12, where lower feature values are predominantly associated with the HVO100 (the "True" label), confirming the consistency between the SHAP values and KDE visualizations. These findings highlight the direct relationship between the statistical significance of frequency domain features (as reflected by low p-values) and their practical effectiveness in distinguishing between classes (as illustrated by the KDE plots). Such insights confirm the effectiveness in the current feature engineering approach using tsfresh and provide reference for future improvement, which will be discussed more in detail in Section 4.3.

4.3. Frequency domain analysis

It is our hypothesis that this is caused by the fuel properties and combustion characteristics. The above evidence indicated that the Fourier transform and wavelet transform of the original signal (rotational speed) could contribute significantly to distinguish the fuel types. The superior effectiveness of frequency domain features in the models can be attributed to several factors: the inherent physical and chemical mechanism differences between HVO100 and EN590 fuels, such as ignition delay and heating value, influence the engine's dynamic response and result in subtle variations in periodicity, harmonics, and transient behaviors of the engine's rotational speed signal. While time-domain features like mean, variance, and skewness capture overall trends and simple fluctuations, they often lack the sensitivity to detect the nuanced, periodic patterns or oscillations introduced by different fuel types. In contrast, frequency domain features, extracted through Fourier or wavelet transforms, can reveal these hidden periodicities and transient events by decomposing the signal into its constituent frequencies. Additionally, frequency domain analysis aids in separating signal components related to fuel-specific combustion phenomena from background noise or unrelated engine operations, thereby enhancing the signal-to-noise ratio for the most discriminative features. The current results, including SHAP and p-value analyses, consistently highlighted frequency domain features (such as specific Fourier and wavelet coefficients) as having the highest importance in the model, indicating that these features more effectively encode combustion-induced differences and are thus crucial for distinguishing between HVO100 and EN590 compared to time-domain statistics.

As a result, we will first explain our motivation for this hypothesis by

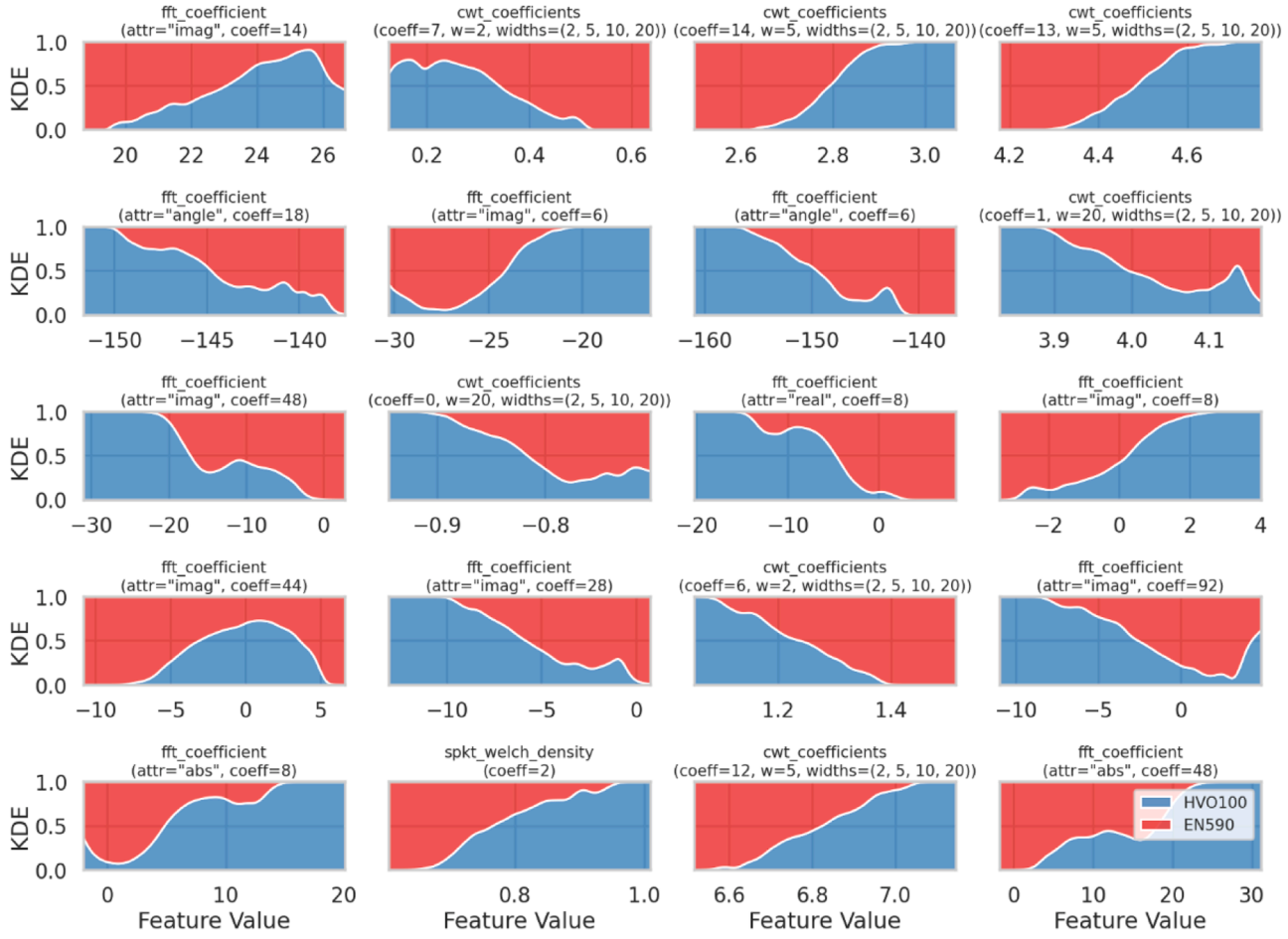

Fig. 13. Normalized kernel density estimate (KDE) plots of the top 20 features used for AutoML experiments at 10,000 Hz based on p-values (cf. Fig. 7).

discussing the difference between two fuels in Section 4.3.1, then discuss the implications of Fourier transform (cf. Section 4.3.2) and wavelet transform (cf. Section 4.3.3).

4.3.1. Fuel properties and combustion characteristics

Despite that HVO100 and EN590 share a lot of similarities, their cetane numbers are significantly different. Cetane number is a measure of the fuel's ignition quality, EN590 typically has a cetane number between 51 and 56, while HVO100 has a higher cetane number, usually above 70 [8,50]. Specifically in this research, NESTE measured the cetane number for both of them, it is 51.8 for EN590 and 72.5 for HVO100. A higher cetane number results in shorter ignition delay time. In addition, depending on the engine calibration, a higher cetane number also lead to other different combustion characteristics [9].

During one cycle of engine combustion, ignition is a relatively short process compared to the overall combustion. Assuming that ignition delay characteristics manifest as a transient signal, and some other combustion characteristics are reflected as one or more stationary signal (s), it is plausible that applying Fourier and wavelet transforms to the engine's rotational speed signal could help distinguish fuel types. Fourier transform would capture the stationary components, while wavelet transform could identify both stationary and transient aspects (cf. Section 3.4.2 and 3.4.3). This forms our hypothesis to explain why fft_coefficient and cwt_coefficient were previously identified as most relevant features and could potentially provide insights into the different combustion characteristics of these fuels. However, tsfresh at its efficient configuration (cf. Section 3.2.1) by default only extracts a limited range from the results of Fourier and wavelet transform of the data. To remedy this problem, Section 4.3.2 and 4.3.3 employ a wider range analysis, which includes but is not limited to the ones extracted by the default settings of tsfresh.

4.3.2. Fourier transform

The previously mentioned fft_coefficient (cf. Section 4.2) represents the different components of discrete Fourier transform with the efficient Fast Fourier Transform (FFT) algorithm [33]. The Fourier transform of four selected samples (same as in Fig. 3) with downsampling frequency of 10,000 Hz are plotted in Fig. 14 to give readers an intuitive understanding. This figure displays the amplitudes and phases of various frequencies, where the different colors represent different fuel types. Since the SHAP values and p-values in Section 4.2 highlighted the relevance of the 6th, 8th and 48th Fourier coefficients, their corresponding frequencies are marked with dashed lines, namely 37.5, 50, and 300 Hz, according to Eq. (6). While Fig. 14 shows slight variations in both amplitude and phase, the differences between the samples of two different fuels are not obvious. Moreover, it does not clearly indicate that the critical frequencies of 37.5, 50, and 300 Hz effectively differentiate the classes.

To further analyze the components of the Fourier transform at 37.5, 50, and 300 Hz, KDE and polar plots are presented in Fig. 15 using all the samples. In these polar plots, the distance from the center represents the component's normalized amplitude, while the angle indicates the phase. Different colors are used to distinguish between the classes. For a fixed frequency, the difference between the scatter polar plots of the two fuel types are very subtle but aligns with the KDE plots. For example, a greater real part of the 50 Hz Fourier component (i.e. coefficient number 8) indicates EN590, and lesser imaginary part of the 300 Hz Fourier component (i.e. coefficient number 48) indicates HVO100. Though, no boundary between the fuel types can be found by analyzing the components one by one.

Additionally, it should be noted that a wider range of components of Fourier transform beyond the default settings in tsfresh were also extracted and analyzed. However, results have been omitted since these components did not yield any new insights.

In general, when the Fourier transform is applied to the original signal (rotational speed), it can reveal the stationary component of the signal. This can be useful in distinguishing between different fuels. However, the difference between fuels may not be easily noticeable through Fourier transform alone. Therefore, it is beneficial to combine machine learning models with Fourier transform to effectively classify different fuels, as suggested by the proposed method.

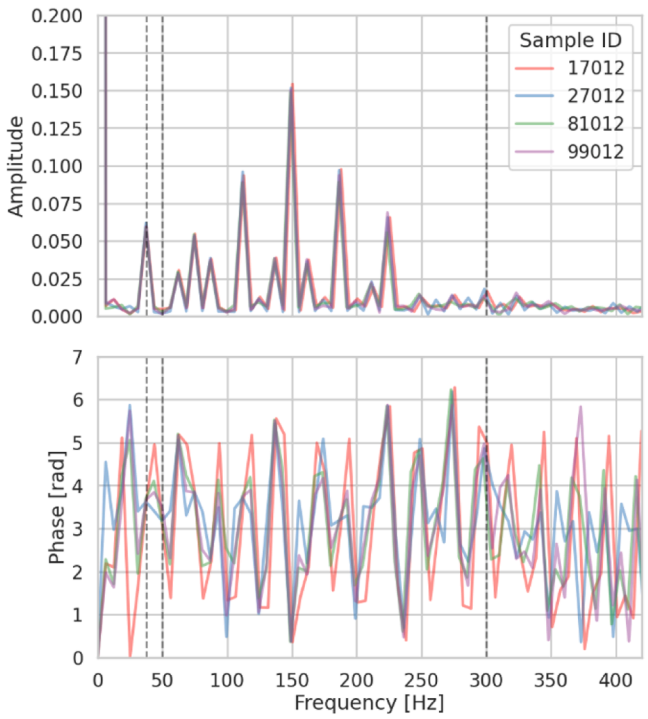

Fig. 14. Fourier transform of four samples as in Fig. 3. HVO100 was used for samples with ID 17,012 and 27,012, and EN590 was used for samples with ID 81,012 and 99,012. The vertical lines highlight 37.5, 50 and 300 Hz, which correspond to the 6th, 8th and 48th Fourier coefficients, respectively.

4.3.3. Wavelet transform

Previous results indicate that cwt_coefficients (i.e. components from wavelet transform) can capture important differences between the two fuel types in terms of time dependent frequency components. In further analysis, the parameter space was increased compared to the original one used by default in tsfresh. To generate the scales a_n , each element in [2, 5, 10] was multiplied with each element in [10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}], resulting in 18 ($= 3 \times 6$) scales from 0.0002 up to 100. For the positions b_n, a linear range from 0 to 900 was chosen, with increments of 10, corresponding to time points from 0 to 0.09 s at a time step $\Delta t = 0.001$ s. This ensured coverage of at least one full revolution, approximately 800 sampling points (0.08 s).

The most important findings were several cwt_coefficients with $a_n = 50$ and b_n around 600 (0.06 s) giving p-value of 0, indicating highly relevant features. This is displayed in Fig. 16, where KDE plots for these components and different choices of coefficient value are shown. As opposed to Fourier signals which are stationary, this component has an apparent transient behavior. One can see that the difference between two fuels is very pronounced around coefficient of 600 (the middle row of Fig. 16), whereas their difference becomes less distinctive at other coefficients (the first and last row of Fig. 16). These results suggest a characteristic difference between the two fuel types at 0.06 s at scale of 50, which could be the overall representation of different combustion characteristics, including but not limited to ignition delay, that occurred throughout the combustion process.

The application of wavelet transform appears to be a promising method for capturing significant time-frequency characteristics of signals. By analyzing the values of various cwt_coefficients, it is possible to

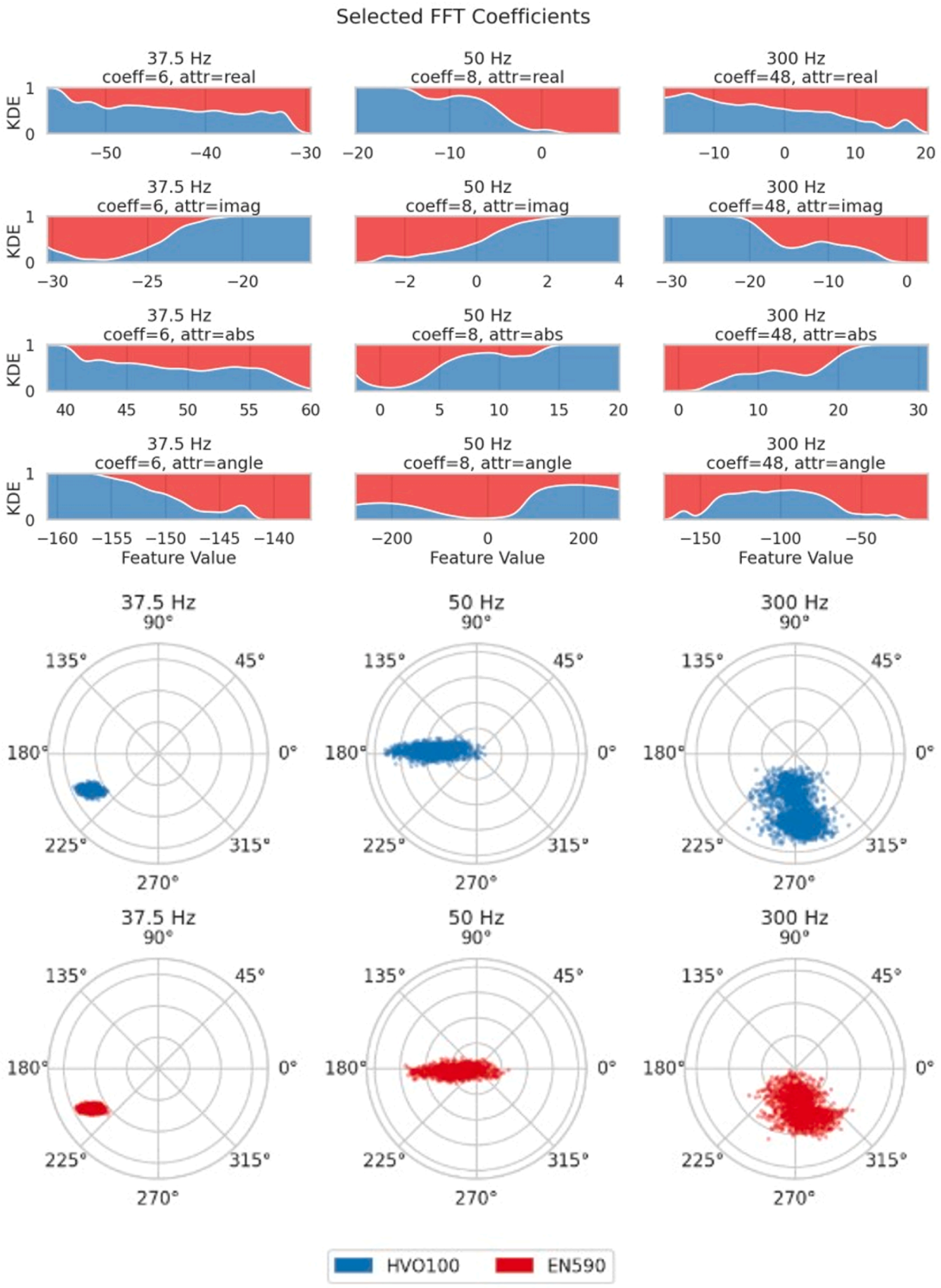

Fig. 15. Normalized KDE plots (top) and polar plots (bottom, amplitudes and phases components) of Fourier transform at 37.5, 50 and 300 Hz for all samples.

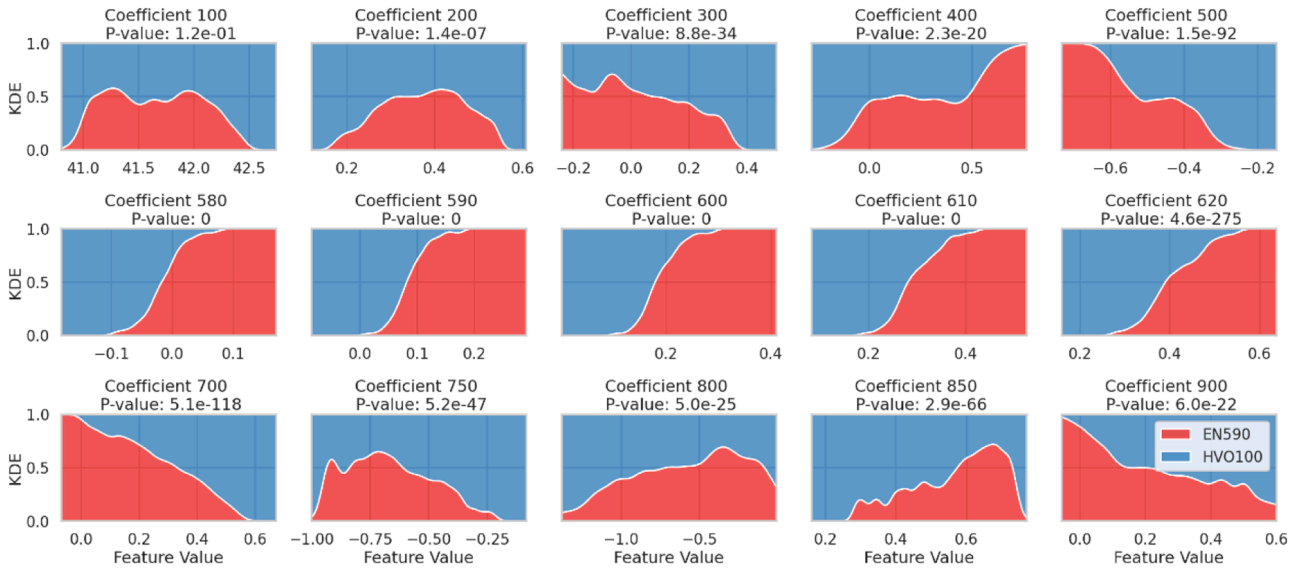

Fig. 16. Multiple filled of normalized kernel density estimate (KDE) plots of components of wavelet transform of all samples of rotational speed at the scale of 50.

differentiate between different fuel types. However, when observing the cwt_coefficients individually and independently, the complex and subtle differences between the signals of different fuel types may not be readily apparent. To overcome this limitation, combining the cwt_coefficients with other types of features can be beneficial. By incorporating additional features and utilizing machine learning methods, it becomes possible to successfully classify the different fuel types. This approach allows for a more comprehensive analysis of the signals, considering multiple aspects and enhancing the accuracy of the classification process.

This is consistent with the findings in Section 4.3.2 that incorporating machine learning alongside frequency domain analysis is beneficial. Solely relying on frequency domain analysis may prove challenging in achieving the desired goal. Therefore, it is necessary to combine machine learning techniques with frequency domain analysis to enhance the effectiveness of the analysis and improve the chances of achieving the desired outcome. By leveraging the power of machine learning algorithms, we can leverage the strengths of both approaches and overcome the limitations of relying solely on frequency domain analysis. Some suggestions are proposed in the 4th point in Section 6.

5. Conclusion

This works aims to predict fuel type based rotational speed data collected from the sensors of Volvo Engine D13K540, EU6SCR. The data was first downsampled to 100, 1000 and 10,000 Hz, then feature engineered using time series feature extraction based on hypothesis tests (tsfresh), then trained through Databricks' AutoML.

Under the configured conditions, the fuel type in the engine could be predicted using the engine rotational speed. The study yields robust results and could achieve 0.995 as the best test F1 score. With minor exceptions, better test F1 score could be achieved with more features and, especially, higher downsampling frequency. However, the marginal benefit might be higher to add 5 more features instead of 10 times the downsampling frequency, especially in resource-constrained environments.

Certain features such as fft_coefficient(attr = "img", coeff = 6) and cwt_coefficient(coeff = 14, w = 5, width = (2, 5, 10, 20)) consistently display high mean absolute SHAP values and low p-values across AutoML experiment runs, making them the leading indicators to differentiate between two fuel types.

Applying Fourier transform and wavelet transform to the rotational speed could capture the stationary and transient components of the signal in the frequency domain, which exhibit differences among different fuels. However, the unobvious nature of these differences indicates the benefits of using machine learning models in classifying the fuel type, as we have done in this work.

In summary, this research introduces an uncomplicated, easily understandable, and computationally efficient machine learning method for accurately predicting the type of fuel used in industrial engines. The results of the study highlight the promising prospects of implementing this approach in practical production settings.

6. Future works

Considering the research objectives and limited resources available, there are several aspects that have not been explored in this work and require further attention in future studies.

Firstly, although fft_coefficient and cwt_coefficients have been identified as important features, the default settings of tsfresh only extract a limited range. Similarly, AutoML has only explored a restricted range of hyperparameters. Therefore, expanding the range of feature extraction and hyperparameter tuning could potentially enhance the performance of the model. One good approach is to start with a broad but coarse grid search to identify promising coefficients based on statistical significance (p-value), feature importance (SHAP values), and model performance. Once key coefficients are identified, one could perform a finer search around these candidates and iterate this process until predefined criteria are met. Specifically, in this case, for frequency-related coefficients, low-order coefficients are prioritized, as they capture general trends and are less sensitive to noise. For time-related coefficients, time periods that are physiochemically relevant (such as the ignition phase), which may reflect differences between fuel types, are of interest.

Secondly, a significant portion of the paper focuses on relevant features at a downsampling frequency of 10,000 Hz. It would be beneficial to dedicate more effort to investigate the relevant features at frequencies of 100 Hz and 1000 Hz. In addition, one needs to identify and remove redundant features to optimize model performance and computational efficiency.

Thirdly, the near-perfect test F1 score of 0.995 raises concerns that it is too good to be true. One cause could be that the model was trained solely based on combustion experiments with limited operating conditions. To improve the model's universality, it is necessary to conduct

additional combustion experiments and collect more training data.

Fourthly, the current work can be considered as a form of data mining aimed at narrowing down the investigation scope of the combustion process. The analysis conducted in the frequency domain using Fourier and wavelet transforms has provided insights into a specific time range that requires attention in order to comprehend the differences in the combustion process between the two fuel types. Moving forward, it would be beneficial to explore additional parameters such as pressure, temperature, and species during this identified specific time range. By doing so, one can effectively pinpoint the combustion events that could easily distinguish between these fuels and provide both physical and chemical explanations, rather than relying on mere machine learning and statistical testing.

Fifthly, the current work analyzed data at different experiment conditions together. At the time, the primary goal was to verify whether representative common features in the signals under real-world, noisy, and variable operating conditions were present. The aim was to quickly assess whether potential for broader application existed in this approach. Additionally, it was intended that machine learning techniques would be leveraged for data mining to identify "low-hanging fruit", i.e. features that are easy to extract and representative, which could guide feature selection and model optimization. However, it is apparent that isolating the intrinsic features of the signals improves the interpretability of the analysis and should be further explored.

By addressing these aspects in future work, one could further enhance the model's performance, explore a wider range of features and hyperparameters, and ensure its applicability across various operating conditions, especially in industrial production settings.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used generative AI service provided by Volvo Group in order to improve the readability and language of the manuscript. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT authorship contribution statement

Ning Guo: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Erik Jansson:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Mattias Johansson:** Writing – review & editing, Methodology, Investigation, Formal analysis, Conceptualization. **Ronny Lindgren:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Andreas Nyman:** Writing – review & editing, Supervision, Conceptualization. **Jonas Sjöblom:** Writing – review & editing, Resources, Methodology, Investigation, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Ning Guo, Erik Jansson, Mattias Johansson, Ronny Lindgren, Andreas Nyman, Jonas Sjöblom have patent pending to Volvo Penta Corporation. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work related to data engineering, analytics, and machine learning is funded by AB Volvo Penta via technical development project "Applied AA and AI" (Applied Advanced Analytics and Artificial Intelligence, Project Number 3197). Additionally, the combustion experiment receives funding from the Combustion Engine Center (CERC), "The Area of Advance: Transport" and CHAIR (Chalmers AI Research) at Chalmers University of Technology, Sweden.

Andreas B. Ofner at Know-Center GmbH is highly appreciated in providing background information. Kristian Adelsund from AB Volvo Penta is acknowledged for his input in technical discussions.

Data availability

The authors do not have permission to share data.

References

[1] Manavalan E, Jayakrishna K. A review of Internet of Things (IoT) embedded sustainable supply chain for industry 4.0 requirements. Comput Ind Eng 2019;127:925–53. https://doi.org/10.1016/j.cie.2018.11.030. 2019/01/01/.

[2] Lu Y, Xu X, Wang L. Smart manufacturing process and system automation – A critical review of the standards and envisioned scenarios. J Manuf Syst 2020;56:312–25. https://doi.org/10.1016/j.jmsy.2020.06.010. 2020/07/01/.

[3] Lee J, Bagheri B, Kao H-A. A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems. Manuf Lett 2015;3:18–23. https://doi.org/10.1016/j.mfglet.2014.12.001. /01/01/2015.

[4] Sagin SV, Sagin SS, Fomin O, et al. Use of biofuels in marine diesel engines for sustainable and safe maritime transport. Renew Energy 2024;224:120221. https://doi.org/10.1016/j.renene.2024.120221. /04/01/2024.

[5] Smigins R, Sondors K, Pirs V, Dukulis I, Birzietis G. Studies of engine performance and emissions at full-load mode using HVO, diesel fuel, and HVO5. Energies 2023;16(12). https://doi.org/10.3390/en16124785.

[6] Bullermann J, Meyer N-C, Krafft A, Wirz F. Comparison of fuel properties of alternative drop-in fuels with standard marine diesel and the effects of their blends. Fuel 2024;357:129937. https://doi.org/10.1016/j.fuel.2023.129937. /02/01/2024.

[7] Batista GEAPA, Keogh EJ, Tataw OM, de Souza VMA. CID: an efficient complexity-invariant distance for time series. *Data* Min *Knowl Discov* 2014;28(3):634–69. https://doi.org/10.1007/s10618-013-0312-3. 2014/05/01.

[8] Yadav J, Deppenkemper K, Pischinger S. Impact of renewable fuels on heavy-duty engine performance and emissions. Energy Rep 2023;9:1977–89. https://doi.org/10.1016/j.egyr.2023.01.016. 2023/12/01/.

[9] Han J, Wang Y, Somers LMT, van de Beld B. Ignition and combustion characteristics of hydrotreated pyrolysis oil in a combustion research unit. Fuel 2022;316:123419. https://doi.org/10.1016/j.fuel.2022.123419. /05/15/2022.

[10] So Khuong L, Hashimoto N, Konno Y, Suganuma Y, Nomura H, Fujita O. Droplet evaporation characteristics of hydrotreated vegetable oil (HVO) under high temperature and pressure conditions. Fuel 2024;368:131604. https://doi.org/10.1016/j.fuel.2024.131604. /07/15/2024.

[11] Aliramezani M, Koch CR, Shahbakhti M. Modeling, diagnostics, optimization, and control of internal combustion engines via modern machine learning techniques: a review and future directions. Prog Energy Combust Sci 2022;88:100967. https://doi.org/10.1016/j.pecs.2021.100967. /01/01/2022.

[12] Wojcieszyk M, Kroyan Y, Kaario O, Larmi M. Prediction of heavy-duty engine performance for renewable fuels based on fuel property characteristics. Energy 2023;285:129494. https://doi.org/10.1016/j.energy.2023.129494. /12/15/2023.

[13] Canal R, Riffel FK, Gracioli G. Machine learning for real-time fuel consumption prediction and driving profile classification based on ECU data. IEEE Access 2024;12:68586–600. https://doi.org/10.1109/ACCESS.2024.3400933.

[14] Patil C, Theotokatos G. Comparative analysis of data-driven models for marine engine In-cylinder pressure prediction. Machines 2023;11(10):926.

[15] Castresana J, Gabiña G, Martin L, Basterretxea A, Uriondo Z. Marine diesel engine ANN modelling with multiple output for complete engine performance map. Fuel 2022;319:123873. https://doi.org/10.1016/j.fuel.2022.123873. /07/01/2022.

[16] Malode SJ, Prabhu KK, Mascarenhas RJ, Shetti NP, Aminabhavi TM. Recent advances and viability in biofuel production. Energy Convers Manag: X 2021;10:100070. https://doi.org/10.1016/j.ecmx.2020.100070. /06/01/2021.

[17] Nath S. Biotechnology and biofuels: paving the way towards a sustainable and equitable energy for the future. Discov Energy 2024;4(1):8. https://doi.org/10.1007/s43937-024-00032-w. /06/14 2024.

[18] Karim F, Majumdar S, Darabi H, Harford S. Multivariate LSTM-FCNs for time series classification. Neural Netw 2019;116:237–45. https://doi.org/10.1016/j.neunet.2019.04.014. 2019/08/01/.

[19] Zhao J, shapeDTW Itti L. Shape dynamic time warping. Pattern Recognit 2018;74:171–84. https://doi.org/10.1016/j.patcog.2017.09.020. /02/01/2018.

[20] Hong F, Chen J, Zhang Z, Wang R, Gao M. Time series risk prediction based on LSTM and a variant DTW algorithm: application of bed inventory overturn prevention in a pant-leg CFB boiler. IEEE Access 2020;8:156634–44. https://doi.org/10.1109/ACCESS.2020.3009679.

[21] Choi K, Yi J, Park C, Yoon S. Deep learning for anomaly detection in time-series data: review, analysis, and guidelines. IEEE Access 2021;9:120043–65. https://doi.org/10.1109/ACCESS.2021.3107975.

[22] Christ M, Braun N, Neuffer J, Kempa-Liehr AW. Time series FeatuRe extraction on basis of scalable hypothesis tests (tsfresh – A Python package). Neurocomputing 2018;307:72–7. https://doi.org/10.1016/j.neucom.2018.03.067. /09/13/2018.
[23] FACT SHEET engine D13K540, EU6SCR. PDF document, [Online]. Available: https://stpi.it.volvo.com/STPIFiles/Volvo/FactSheet/D13K540,%20EU6SCR_Eng_07_310999624.pdf Accessed: 2024/09/10.
[24] Middlehurst M, Schäfer P, Bagnall A. Bake off redux: a review and experimental evaluation of recent time series classification algorithms. *Data* Min *Knowl Discov* 2024. https://doi.org/10.1007/s10618-024-01022-1. 2024/04/19.
[25] Abanda A, Mori U, Lozano JA. A review on distance based time series classification. *Data* Min *Knowl Discov* 2019;33(2):378–412. https://doi.org/10.1007/s10618-018-0596-4. 2019/03/01.
[26] Ye L, Keogh E. Time series shapelets: a novel technique that allows accurate, interpretable and fast classification. *Data* Min *Knowl Discov* 2011;22(1):149–82. https://doi.org/10.1007/s10618-010-0179-5. 2011/01/01.
[27] Foumani NM, Miller L, Tan CW, Webb GI, Forestier G, Salehi M. Deep Learning for time series classification and extrinsic regression: a current survey. ACM Comput Surv 2024;56(9). https://doi.org/10.1145/3649448. Article 217.
[28] Lubba CH, Sethi SS, Knaute P, Schultz SR, Fulcher BD, Jones NS. catch22: cAnonical time-series CHaracteristics. *Data* Min *Knowl Discov* 2019;33(6):1821–52. https://doi.org/10.1007/s10618-019-00647-x. 2019/11/01.
[29] Christ M., Kempa-Liehr A.W., Feindt M. Distributed and parallel time series feature extraction for industrial big data applications. *arXiv preprint* arXiv:161007717. 2016;doi:https://doi.org/10.48550/arXiv.1610.07717.
[30] Zhao Y, Meng X, Qi T, et al. AI-based rainfall prediction model for debris flows. Eng Geol 2022;296:106456. https://doi.org/10.1016/j.enggeo.2021.106456. /01/01/2022.
[31] Petelin G, Cenikj G, Eftimov T. Towards understanding the importance of time-series features in automated algorithm performance prediction. Expert Syst Appl 2023;213:119023. https://doi.org/10.1016/j.eswa.2022.119023. /03/01/2023.
[32] Long T, Outerleys J, Yeung T, et al. Predicting ankle and knee sagittal kinematics and kinetics using an ankle-mounted inertial sensor. Comput Methods Biomech Biomed Engin 2024;27(9):1057–70. https://doi.org/10.1080/10255842.2023.2224912. 2024/07/03.
[33] Cooley JW, Tukey JW. An algorithm for the machine calculation of complex fourier series. Math Comput 1965;19(90):297–301. https://doi.org/10.2307/2003354.
[34] Welch P. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans Audio Electroacoust 1967;15(2):70–3. https://doi.org/10.1109/TAU.1967.1161901.
[35] Box GE, Jenkins GM, Reinsel GC, Ljung GM. Time series analysis: forecasting and control. John Wiley & Sons; 2015.
[36] Bandt C, Pompe B. Permutation entropy: a natural complexity measure for time series. Phys Rev Lett 2002;88(17):174102. https://doi.org/10.1103/PhysRevLett.88.174102. 04/11/.
[37] Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 1947;18(1):50–60.
[38] Fay MP, Proschan MA. Wilcoxon-Mann-Whitney or *t*-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. Stat Surv 2010;4:1. https://doi.org/10.1214/09-SS051.
[39] Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat 2001;29(4):1165–88. 24.
[40] He X, Zhao K, Chu X. AutoML: a survey of the state-of-the-art. Knowl Based Syst 2021;212:106622. https://doi.org/10.1016/j.knosys.2020.106622. 2021/01/05/.
[41] Databricks L'Esteve R. editor. In: L'Esteve R, editor. The azure data lakehouse toolkit: building and scaling data lakehouses on azure with delta lake, apache spark, databricks, synapse analytics, and snowflake. Apress; 2022. p. 83–139.
[42] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J mach Learn res 2011;12:2825–30.
[43] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. presented at. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. https://doi.org/10.1145/2939672.2939785.
[44] Ke G, Meng Q, Finley T, et al. Lightgbm: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 2017;30.
[45] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30.
[46] Mallat S. A wavelet tour of signal processing. Elsevier; 1999.
[47] Ricker N. Further developments in the wavelet theory of seismogram structure*. Bull Seismol Soc Am 1943;33(3):197–228. https://doi.org/10.1785/bssa0330030197.
[48] Ricker N. Wavelet functions and their polynomials. GEOPHYSICS 1944;9(3):314–23. https://doi.org/10.1190/1.1445082.
[49] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genom 2020;21(1):6. https://doi.org/10.1186/s12864-019-6413-7. 2020/01/02.
[50] Dimitriadis A, Natsios I, Dimaratos A, et al. Evaluation of a hydrotreated vegetable oil (HVO) and effects on emissions of a passenger car diesel engine. Original research. Front Mech Eng 2018:4. https://doi.org/10.3389/fmech.2018.00007. 2018-July-31.