

Article

Diesel Adulteration Detection with a Machine Learning-Enhanced Laser Sensor Approach

Bachar Mourched *, Tariq AlZoubi  and Sabahudin Vrtagic 

College of Engineering and Technology, American University of the Middle East, Egaila 54200, Kuwait; tariq.alzoubi@aum.edu.kw (T.A.); sabahudin.vrtagic@aum.edu.kw (S.V.)

* Correspondence: bachar.mourched@aum.edu.kw

Abstract: This paper introduces a novel and cost-effective method for detecting adulterated diesel, specifically targeting contamination with kerosene, by leveraging machine learning and the refractive index values of mixed diesel samples. It proposes a laser-based sensor, employing COMSOL simulations for synthetic data generation to facilitate machine learning training. This innovative approach not only streamlines the detection process by eliminating the need for expensive equipment and specialized personnel but also enables on-site testing without extensive sample preparation. The sensor's design, utilizing light refraction and reflection principles, allows for the accurate measurement of diesel adulteration levels. Validation results showcase the machine learning models' high precision in predicting adulteration percentages, as evidenced by an R-squared value of 0.999 and a mean absolute error of 0.074. This research signifies a leap in sensor technology, offering a practical solution for rapid diesel adulteration detection, especially in developing countries, by minimizing reliance on advanced laboratory analyses. The sensor's design aligns with the requirements for low-cost IoT technology, presenting a versatile tool for various applications.

Keywords: light reflection/refraction; sensor; refractive index; diesel adulteration; kerosene; COMSOL Multiphysics; machine learning; models



Citation: Mourched, B.; AlZoubi, T.; Vrtagic, S. Diesel Adulteration Detection with a Machine Learning-Enhanced Laser Sensor Approach. *Processes* **2024**, *12*, 798. <https://doi.org/10.3390/pr12040798>

Academic Editor: Jiaqiang E

Received: 14 March 2024

Revised: 8 April 2024

Accepted: 9 April 2024

Published: 16 April 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Diesel adulteration is a widespread issue in many countries, driven by the price disparity between similar quantities of different products [1]. This practice poses a significant threat to national interests, with petroleum products like high-speed diesel and petrol being particularly vulnerable due to their high demand, cost, and occasional scarcity.

Among these, diesel, extensively used in heavy vehicles, is commonly tampered with using domestically available, subsidized kerosene, causing substantial damage to vehicle engines and reducing fuel efficiency [2]. The similar properties of kerosene and diesel facilitate this illicit practice for financial gain. Kerosene, with a calorific value of 45 KJ/g, is distributed at reduced rates by some governments for household and industrial use, yet unscrupulous individuals mix it with diesel, exploiting these overlapping properties [3].

Diesel, a complex blend of hydrocarbons (C9–C19) with a specific calorific value and distillation range and a composition that includes 15–30% aromatics and 70–85% saturated aliphatics, contrasts with kerosene's hydrocarbon range (C6–C16) [4]. Despite efforts by government bodies to monitor diesel and petrol quality through random sample collection, existing laboratory analyses are flawed, allowing adulterated samples to pass. The literature lacks fully accurate and cost-effective methods for qualitative adulteration assessment in diesel. Recent advancements claim more precise quantification methods, though some physical parameter-based analyses remain impractical.

Specific gravity is a measure of the density of a substance compared to the density of water at a specified temperature. For diesel, this property is important because it affects the fuel's energy content, combustion characteristics, and behavior in engines. The specific

gravity of diesel samples ranges between 0.800 and 0.850. In comparison, kerosene's specific gravity spans from 0.780 to 0.820, showing a considerable overlap [5]. Therefore, it becomes difficult to distinguish between pure diesel, pure kerosene, and their mixtures based solely on this property. This overlap presents a significant challenge for quality control, requiring more sophisticated analytical techniques to accurately detect and quantify the extent of adulteration.

The process of separating and identifying polycyclic aromatic compounds in diesel can be efficiently conducted using techniques such as two-dimensional microbore high-performance liquid chromatography [6], high-resolution mass spectrometry [7], and micro-fabricated gas chromatography [8], yet challenges persist due to the hydrocarbon overlap in diesel and kerosene, and conclusive detection remains complex.

Jabin et al. [9] introduced a novel approach for detecting diesel adulteration using a silver-coated surface plasmon resonance (SPR)-based biosensor. The sensor's performance was analyzed using COMSOL Multiphysics V-5.1 and MATLAB-V16 software. Their experimental evaluation focused on key optical parameters of the SPR-PCF (photonic crystal fiber), including birefringence, coupling length, power fraction, etc. This advancement is considered significant in the field of photonics.

Another research work evaluated the efficacy of infrared (IR) spectroscopies [10] in conjunction with advanced statistical models like partial least squares (PLS) regression, support vector machine regression (SVR), and multivariate curve resolution with alternating least squares (MCR-ALS) for quantitatively and qualitatively identifying kerosene in commercial diesel. These models proved accurate in quantifying kerosene concentrations ranging from 2.5% to 40% by volume, with low errors ($\text{RMSEC} < 2.59\%$ and $\text{RMSEP} < 5.56\%$) and a high correlation between actual and predicted values.

Nonetheless, additional methods have also been employed in this domain, including NMR spectroscopy [11–13], fiber optic technique [14–16], fluorescent paper strips [17], optical sensor [18,19], infrared spectrometer [20–23], ultrasonic technique [24], artificial intelligence (AI) prediction [25], computational technique [26], and many other chemical-based techniques [27–29].

Recent advancements have positioned machine learning models as a novel approach for examining diesel adulteration. Bhowmik et al. [25] explored how ethanol blended with adulterated diesel could improve exhaust emissions without compromising engine performance. Utilizing experimental findings, a gene expression programming (GEP) model, rooted in artificial intelligence (AI) and encompassing multiple parameters, was crafted to delineate the connection between various inputs (e.g., engine load and shares of kerosene and ethanol) and outputs (e.g., BTE, BSEC, NOx, UHC, and CO) for Diesosenol implementations. This model demonstrated remarkable accuracy, validated by a comparison with empirical data and statistical evaluations, displaying minimal mean square error values between 0.00002 and 0.00031.

Each of these techniques can provide valuable information about the presence of adulterants in diesel. However, the choice of method often depends on the specific requirements of the testing, including the need for accuracy, speed, cost-effectiveness, and the ability to perform analyses on-site. Some of these techniques, such as SPR or IR spectroscopy [9,10], have a high initial cost for equipment and may require calibration for different types of diesel. It also might be less effective in distinguishing between similar types of hydrocarbons or detecting low concentrations of adulterants. Others, such as gas and liquid chromatography [8,27], require skilled operators, time-consuming sample preparation, and are not suitable for on-site testing. The equipment is expensive as well and requires skilled operation. Furthermore, no technique can be suitable for all types of adulterants, particularly organic ones. Chemical sensor techniques may not detect unknown adulterants or those present in very low concentrations, and sensor degradation over time can affect reliability [30].

Machine learning and neural network-based techniques can analyze complex datasets using spectroscopic techniques (like NIR and FTIR) to improve the accuracy and prediction

of adulteration levels. AI can handle large data volumes, making it suitable for detecting subtle patterns indicative of adulteration. Yet, this method requires extensive datasets for training and may be complex to set up, and the accuracy depends significantly on the quality and diversity of the training data [31].

Relative to the previously discussed methods and techniques that necessitate expensive equipment, consumables, laboratory apparatus, and operation by trained professionals, certain optical techniques may prove more economical in the long term. This cost-efficiency stems from their reduced need for consumables, advantages like lasting durability, low upkeep, the possibility for automated operation, and the capability to conduct tests without damaging the samples [18–23].

However, optical techniques also have disadvantages, such as limited sensitivity, especially when certain types of adulterants have similar refractive indices or densities to diesel, which makes the adulterants difficult to detect. They are also not effective for precise quantification of adulteration levels.

The use of optical sensors alongside digital and AI technologies serves as a prospective means of improving the efficiency of the performance and analysis of data; hence, this approach could significantly reduce costs and lead to the development of superior detection methods. Even though it is a yet-understudied method used for detecting diesel fraud, distressingly, few studies have investigated this issue [32–36].

This manuscript introduces a novel method combining light reflection principles with a machine learning ML framework to detect diesel adulteration with kerosene instantly. Traditional methods for detecting fuel adulteration, as aforementioned, often involve complex laboratory analyses, which are time-consuming and not feasible for real-time application. Recent advancements have explored optical techniques and ML algorithms for adulteration detection; however, these approaches have faced challenges, including the need for extensive real-time data for ML model optimization and limitations in sensitivity and specificity.

Our research stands at the forefront by employing COMSOL Multiphysics and ML to design a sensor that overcomes these challenges. Unlike previous studies that separate the application of optical studies and ML, our approach integrates them, enhancing the ability to detect adulteration with high precision even with limited data sources. This integration is crucial, considering the dynamic nature of diesel properties and the variety of adulterants used. The novelty of our study lies in the creation of synthetic data through simulations in COMSOL Multiphysics, enabling the training of an ML model without the extensive need for real-world contaminated samples (reference the importance of synthetic data in overcoming data scarcity in ML models).

Moreover, the proposed model introduces a cost-effective laser configuration, leveraging Snell's law to analyze light interactions with adulterated diesel. This approach allows for the estimation of kerosene concentrations in diesel, a significant step forward in real-time adulteration detection.

In summary, this manuscript contributes a unique perspective to the field by merging optical principles with ML for detecting diesel adulteration. It offers a new, affordable, and portable solution for on-the-spot analysis, addressing a significant gap in current research. Future works will delve into the development phase and experimental validations, further solidifying this innovative approach's applicability and effectiveness.

This paper is structured to outline the conceptual framework and methodology behind the proposed sensor in Section 2, laying the groundwork for the methodologies applied. Section 3 provides a detailed exposition of the machine learning approach, including the definition of the dataset and the models utilized within the machine learning framework. Section 4 then presents the results, along with a discussion evaluating the effectiveness and performance of the machine learning models deployed.

2. Optical-Based Sensor Mechanism

COMSOL Multiphysics® [37] is used to define the sensor's operating principle and generate synthetic data for the input-output of the regressors in the machine learning tool. This simulation entails representing the layout of the sensor as a 2D rectangular configuration, akin to a container or tank in which the diesel sample is subjected to testing.

The fundamental principle underpinning the sensor's functionality involves the refraction and reflection of laser light within the diesel sample. Once the laser light is emitted into the sample, it traverses until it reaches the base of the container, where it reflects off the surface and travels back through the sample before exiting toward a designated area known as the "sensing zone", positioned along the laser's trajectory. Within this zone, the distance d , defined as the light path from its point of entry to its intersection with the sensing zone, is calculated. This distance d depends on various factors, including the light's wavelength, transmission angle, refractive index, and the temperature of the diesel sample.

This section outlines the key parameters and equations used in the COMSOL Multiphysics simulation to analyze the sensor's design and operational principles. The sensor model is divided into three primary components: a cap that accommodates the laser source, a container for the diesel sample, and a sensing zone equipped with sensors located on the container's upper surface. Defined within COMSOL, the sensor's geometry is conceptualized as a 2D structure, distinguishing two distinct areas: air (with a refractive index $n_{air} = 1$) and the diesel sample (with a refractive index n_d). Figure 1 displays a three-dimensional depiction of the sensor's conceptual design.

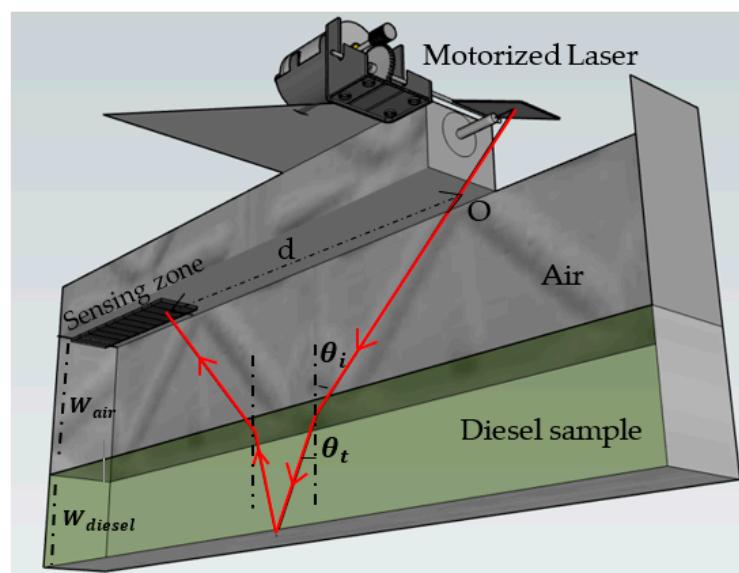


Figure 1. Visual representation displaying the 3D design of the proposed sensor, highlighting the light path from the laser source towards the sensing zone.

Nonetheless, the sensor conceptualization and the data collected via COMSOL in this study pertain to a 2D analysis. The sensor possesses a rectangular geometry, measuring 30 cm in length and 15 cm in width (Figure 2). The laser is positioned at the incidence point O within the cap and emits light at an incidence angle θ_i . This angle can be altered through a system combining a mirror and a servo motor, as illustrated in Figure 1. Upon transitioning from air into a diesel sample, the light beam refracts at an angle θ_t relative to the normal. Snell's law, which establishes the correlation between n_{air} , n_d , θ_i , and θ_t , governs this behavior and is incorporated into the simulation, as described by Equation (1):

$$n_{air} \sin \theta_i = n_d \sin \theta_t \quad (1)$$

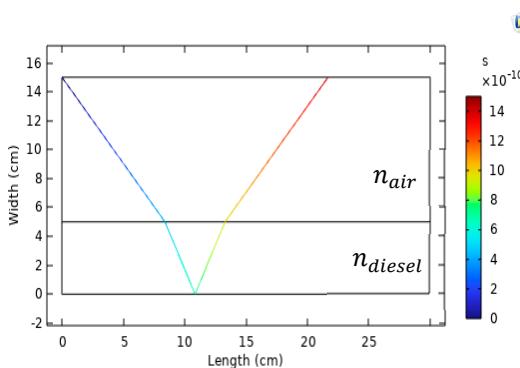


Figure 2. Two-dimensional sensor design in COMSOL Multiphysics.

To reach the sensing zone located a distance d away from the initial point of incidence O, the transmitted light navigates through the diesel sample medium at a refraction angle θ_t . The determination of d is subject to several influencing factors, including the incident light wavelength λ , the angle of incidence θ_i , the refractive index n_d of the diesel sample, and the depth ratio of air to diesel within the medium, denoted as W_{air}/W_{diesel} .

In COMSOL, the modeling of electromagnetic wave propagation utilizes the “Geometrical Optics” time-dependent physics interface. In this study, diffraction effects at the edges and corners of the geometry are neglected by setting the wall boundary conditions to “disappear” options, ensuring perfect absorption. To accurately capture the refraction between air and diesel, the step size for the optical path length is set to 0.01 cm.

COMSOL calculates the ray’s time (source-to-target) to reach the sensor zone. On the other hand, the beam travels at the speed of light, and the duration is expressed in nanoseconds, or Angstroms. It is challenging to locate a time sensor that can precisely measure the nanoscale time difference that occurs between beams after they arrive at the detecting zone. Consequently, it is useless to examine the source-to-target time that COMSOL collected. Thus, the focus of this study is primarily on the distance d parameter, dependent on the λ , n_d , T , θ_i , and W_{air}/W_{diesel} parameters. Simulations are carried out to compute the distance d for a wavelength λ of 450 nm, an angle θ_i ranging from 10° to 40° with a 1° step, a width W_{diesel} ranging from 1 cm to 5 cm with a 0.5 cm step, and a refractive index n_d with a 1×10^{-4} step. The small increment in the n_d computation is designed to guarantee precise identification of the sample’s refractive index, given the close similarity between the refractive indices of diesel and its adulterants.

Numerous studies have explored diesel adulteration by assessing the refractive index of various adulterated diesel samples. Bhausahab et al. [38] tested ten fuel and kerosene samples, each obtained from different reputable sources. Then, different ratios were combined to create admixtures of kerosene in diesel, corresponding to adulteration volume percentages varying from 10% to 100%. Refractive index readings at room temperature from the refractometer for the 10 distinct diesel samples varied from 1.4600 to 1.4612, with an average of 1.4606, and from 1.4445 to 1.4471, with an average of 1.453, for the kerosene samples. For the admixture samples, results show a refractive index of 1.4587, 1.4571, 1.4556, 1.4550, 1.4523, 1.4507, 1.4491, 1.4477, 1.4461, and 1.4444, corresponding to 10% to 100% kerosene adulteration, respectively. It was observed that the refractive index decreases as the proportion of kerosene increases, displaying a linear relationship between the refractive index and the percentage of kerosene.

Kanyathare et al. [39] introduced a method for detecting adulterated diesel oil by comparing the refractive index of mixtures of suspected adulterated and authentic diesel oils using a refractometer. The process benefits from the availability of genuine diesel from regulatory authorities and employs the Lorentz–Lorenz formula to estimate the permittivity changes, aiding in the detection of counterfeit diesel. It suggests the potential for creating a calibration curve library for all diesel types in a country to facilitate screening. The values of the refractive indices obtained were also in the range of the ones obtained in reference [38].

Thus, based on the above, the refractive index is swept in the simulation study from 1.4444 to 1.4604 with 1×10^{-4} increments to cover the maximum range of adulterated diesel concentration.

As aforementioned, the output of the simulation study is the measurement of the parameter d for each change of the parameters θ_i (from 10° to 40° with a 1° increment), W_{diesel} (from 1 cm to 5 cm with a 0.5 cm increment), and n_d (from 1.4444 to 1.4604 with a 1×10^{-4} increment).

In the following section, we will outline the preparation of the dataset by establishing the input-output parameters for the machine learning model and subsequently assess its performance.

3. Machine Learning Regression Models for Diesel Purity Prediction

As referenced above, the simulation COMSOL Multiphysics software is used to analyze the effects of reflection/refraction, which are characterized by multiple variables, including W_{diesel} , θ_i , n_d , and d . Changing these variables will impact the distance d . The effects of different values of each variable are modeled to get a more accurate result. The final outcome is the distance d produced for each variable. The simulation and training variables forming the dataset are:

- Incident angle θ_i from 10° to 40° with a 1° increment.
- Diesel sample depth W_{diesel} from 1 cm to 5 cm with a 0.5 cm increment.
- Refractive index of the diesel sample n_d from 1.4444 to 1.4604 with a 1×10^{-4} increment to cover all possible adulterated diesel volume percentage.

The categorization of diesel adulteration volume percentage is determined by the refractive index, as indicated in Table 1.

Table 1. Refractive index range and corresponding diesel volume percentage adulteration.

n_d Range	Diesel Volume Percentage Adulteration
1.4604 to 1.4588	0 (pure diesel)
1.4587 to 1.4572	10
1.4571 to 1.4557	20
1.4556 to 1.4541	30
1.4540 to 1.4524	40
1.4523 to 1.4508	50
1.4507 to 1.4492	60
1.4491 to 1.4478	70
1.4477 to 1.4462	80
1.4461 to 1.4445	90
≤ 1.4444	100 (pure kerozene)

The data obtained from these simulations were aggregated and reformatted to create a dataset that defines the adulterated diesel volume concentration as output based on the value of the predicted n_d (Table 2). Table 2 presents a portion of the dataset obtained from COMSOL, featuring selected values of incident angles and refractive indices of adulterated diesel at various diesel depths.

The adjusted variables will together create the input dataset for the regression models, which will then be utilized to calculate the percentage of adulteration in diesel. In this section, we delve into the presentation of machine learning regression models to predict diesel concentration/purity based on the data influenced by the light reflection/refraction concept, given in Table 1. We will discuss the pivotal role of normalization, elucidate the process of data partitioning, describe the input–output relationships, introduce relevant equations, and underscore the significance of model evaluation.

Table 2. Input–output data parameters from a subset of the entire dataset.

θi°	W (cm)	Input									Output	
		1	1.5	2	2.5	3	3.5	4	4.5	5	nd	Adulterated Diesel %
10		5.17935546	5.12412847	5.06890149	5.01367451	4.95844752	4.90322054	4.84799356	4.79276657	4.73753959	1.4444	100
		5.17854154	5.1229076	5.06727366	5.01163972	4.95600578	4.90037184	4.8447379	4.78910396	4.73347002	1.4492	60
		5.17781708	5.1218209	5.06582473	5.00982856	4.95383238	4.89783621	4.84184004	4.78584387	4.72984769	1.4535	40
		5.17666367	5.12009079	5.06351791	5.00694504	4.95037216	4.89379928	4.83722641	4.78065353	4.72408065	1.4604	0
15		7.86645862	7.78045004	7.69444146	7.60843288	7.5224243	7.43641572	7.35040714	7.26439856	7.17838999	1.4459	90
		7.8657839	7.77943797	7.69309203	7.6067461	7.52040016	7.43405423	7.34770829	7.26136236	7.17501642	1.4485	70
		7.86408253	7.77688591	7.15694609	7.07622259	6.99549909	6.91477559	6.8340521	6.7533286	6.6726051	1.4551	30
		d (cm)	7.86336562	7.77581054	7.15561411	7.07455762	6.99350112	6.91244463	6.83138814	6.75033164	6.66927515	1.4579
28		15.5735981	15.3847557	15.1959133	15.0070709	14.8182285	14.6293861	14.4405437	14.2517013	14.0628588	1.4475	80
		15.5712769	15.3812739	15.1912709	15.0012679	14.8112649	14.6212619	14.4312589	14.2412559	14.0512529	1.4519	50
		15.5688679	15.3776603	15.1864528	14.9952452	14.8040377	14.6128301	14.4216226	14.230415	14.0392075	1.4565	20
		15.5680341	15.3764097	15.1847853	14.9931608	14.8015364	14.609912	14.4182876	14.2266631	14.0350387	1.4581	10
40		24.4876426	24.1449694	23.8022962	23.459623	23.1169498	22.7742766	22.4316035	22.0889303	21.7462571	1.4456	90
		24.4834672	24.1387064	23.7939455	23.4491846	23.1044238	22.7596629	22.414902	22.0701412	21.7253803	1.4505	60
		24.4812673	24.1354065	23.7895457	23.4436849	23.0978241	22.7519634	22.4061026	22.0602418	21.714381	1.4531	40
		24.4787423	24.131619	23.7844957	23.4373724	23.0902491	22.7431258	22.3960025	22.0488792	21.7017559	1.4561	20

3.1. Dataset Preparation and Partitioning

In our study, we began with a comprehensive dataset comprising 4986 input parameters, encompassing θ_i in degrees, W_{diesel} in cm, n_d and d in cm. To facilitate robust model training and evaluation, we adopted a principled approach to data partitioning. Unambiguously, we allocated 70% of the dataset, amounting to 3490 samples, for model training. The remaining 30%, consisting of 1496 samples, was reserved for rigorous testing of the trained models. Additionally, we generated five sets of data for post-modeling analysis and result presentation, ensuring comprehensive evaluation and validation.

3.2. Normalization for Enhanced Model Performance

Normalization serves as a critical preprocessing step to standardize the input data and ensure uniform scaling across features. Leveraging the “`sklearn.preprocessing.normalize`” function, we transformed θ_i and d data to a consistent range, promoting convergence and stability in regression models. The normalization equation is presented in Equation (2):

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (2)$$

where x represents the input feature, $\min(x)$ is the minimum value of x , and $\max(x)$ is the maximum value of x .

Our input parameters, denoted as $x = [\theta_i, d_1, d_2, \dots, d_7]$, encapsulate the incident angles θ_i and distances d influenced by reflection/refraction and the diesel sample W_{diesel} , while the output variable y represents the refractive index n_d predicted by the regression models.

Regression models are fundamental tools in machine learning for predicting continuous target variables based on input features. In our scenario of predicting diesel purity, regression models play a crucial role in deciphering the complex relationships between incident angles, distances affected by reflection/refraction, and refractive indices. In this study, we explore a collection of regression techniques, including linear regression equations, gradient boosting regressors, decision tree regressors, random forest regressors, extra trees regressors, and voting regressors. Each model encapsulates distinct methodologies to infer the intricate relationships between input features and refractive indices, culminating in accurate predictions of diesel purity. A brief description of each of the mentioned models is presented in the following paragraph.

Linear regression establishes a linear relationship between the input features and the target variable by fitting a straight line to the data points. This model learns the coefficients (slope) and the intercept (bias) that minimize the difference between the predicted and actual values. In our context, linear regression estimates the refractive index based on angles and distances, offering interpretability and simplicity in model representation.

Gradient boosting is an ensemble learning technique that sequentially builds a series of decision trees, each correcting the errors of its predecessor. In our scenario, `GradientBoostingRegressor` constructs a strong predictive model by iteratively minimizing the residuals between predicted and actual refractive indices. By combining weak learners into a robust ensemble, `GradientBoostingRegressor` adapts to complex data patterns and offers superior predictive performance.

Decision trees partition the feature space into hierarchical structures based on feature thresholds, enabling intuitive decision-making. `DecisionTreeRegressor` constructs a binary tree where each node represents a feature and each branch represents a decision based on that feature. In predicting diesel purity, `DecisionTreeRegressor` recursively splits the data to minimize the variance of refractive index predictions, offering transparency and interpretability in model insights.

Random forests leverage the power of ensemble learning by constructing multiple decision trees and aggregating their predictions. `RandomForestRegressor` introduces randomness in tree construction by bootstrapping samples and selecting random subsets of features, thereby reducing overfitting and improving generalization performance. In our

context, RandomForestRegressor captures intricate relationships between angles, distances, and refractive indices, offering robust predictions for fuel purity.

ExtraTreesRegressor, a variant of random forests, introduces additional randomness during tree construction by selecting random thresholds for feature splitting. By incorporating feature randomness and bootstrap sampling, ExtraTreesRegressor explores the feature space more comprehensively, thereby enhancing predictive performance and mitigating overfitting concerns. In predicting diesel purity, ExtraTreesRegressor offers versatility and robustness in capturing subtle variations in angle-distance-refraction relationships.

VotingRegressor aggregates predictions from multiple base estimators, including linear regression, GradientBoostingRegressor, DecisionTreeRegressor, RandomForestRegressor, and ExtraTreesRegressor. By combining diverse regression models, VotingRegressor harnesses the collective wisdom of individual estimators to improve prediction accuracy and robustness. In our scenario, VotingRegressor provides a unified approach to diesel purity estimation, leveraging the strengths of different regression techniques to enhance predictive performance.

Scikit-learn, a popular machine learning library, offers default values for hyperparameters in its implementations of various models. For instance, in gradient boosting, the default learning rate (eta) is typically set to 0.1, while the number of trees ($n_{estimators}$) defaults to 100. Decision trees often have default values such as 'None' for maximum depth (allowing nodes to expand until all leaves are pure) and 2 for the minimum number of samples required to split an internal node (min_samples_split). Random forests usually default to 100 trees ($n_{estimators}$), and while the maximum depth defaults to 'None', for our modeling, it was set to 100. Additionally, 'auto' is the default value for maximum features, which chooses the square root of the number of features for classification and the number of features for regression. Linear regression in scikit-learn does not have hyperparameters to tune by default, but if regularization is applied, the default alpha for Lasso or Ridge regularization is set to 1.0. Similarly, for the extra trees regressor, the default number of trees ($n_{estimators}$) is 100, and while the maximum depth defaults to 'None', it is updated to 100 for this study. The maximum feature defaults to 'auto'. These default values offer a starting point for model training and can be adjusted through hyperparameter tuning to optimize performance for specific datasets and tasks.

To measure the efficacy of our regression models, we employed a suite of evaluation metrics, including mean squared error (MSE), R-squared (R^2), mean absolute error (MAE), and root mean squared error (RMSE). These metrics provide quantitative insights into model performance, enabling nuanced comparisons and informed decision-making. The equations for these metrics are provided in Equations (3)–(6):

$$MSE = \left(\frac{1}{N} \right) \sum_i (y_i - \hat{y}_i)^2 \quad (3)$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (4)$$

$$MAE = \frac{1}{m} \sum_{k=1}^m |y_k - \hat{y}_k| \quad (5)$$

$$RMSE = \sqrt{MSE} \quad (6)$$

where \hat{y}_k represents predicted values for interval k and y_k represents the real output.

In the following section, we will showcase and assess the outcomes derived from the various models employed.

4. Results and Model Verification

Regression models can serve as powerful tools for predicting diesel purity based on incident angle and distance data. Table 3 displays the outcomes for R^2 , MSE, RMSE, and MAE calculated from the different models utilized. It reveals that the models are highly

reliable, as evidenced by an R^2 value of 0.999 and an MAE of 0.074. This confirms the viability of the proposed models for mapping inputs to outputs.

Table 3. Errors in model regression.

	R^2	MSE	RMSE	MAE
GradientBoostingRegressor	0.9995755423	0.0000000090	0.0000948249	0.0000780582
DecisionTreeRegressor	0.9993493505	0.0000000138	0.0001174028	0.0001106952
RandomForestRegressor	0.9998072755	0.0000000041	0.0000638960	0.0000462553
LinearRegression	0.9976417782	0.0000000500	0.0002235102	0.0001711042
ExtraTreesRegressor	0.9999722953	0.0000000006	0.0000242260	0.0000124726
VotingRegressor	0.9997784577	0.0000000047	0.0000685067	0.0000536201

To determine which model shows the best performance, we typically look at various performance metrics like R^2 , MSE, RMSE, and MAE.

In this case, the model with the highest R^2 value and the lowest values for MSE, RMSE, and MAE would generally be considered the best-performing model.

Based on the provided metrics from Table 2, the ExtraTreesRegressor model shows the highest R^2 value (0.9999722953) and the lowest MSE, RMSE, and MAE among all models. Additionally, the RandomForestRegressor also performs exceptionally well, with a high R-squared value (0.9998072755) and low values for MSE, RMSE, and MAE.

Both ExtraTreesRegressor and RandomForestRegressor are ensemble methods and have likely benefited from their ensemble nature and randomness in the model building process, which can help improve generalization and reduce overfitting. Therefore, based on the provided metrics, ExtraTreesRegressor and RandomForestRegressor appear to show the best performance among the models listed.

VotingRegressor was chosen for this study due to its ability to aggregate predictions from multiple base estimators, including linear regression, GradientBoostingRegressor, DecisionTreeRegressor, RandomForestRegressor, and ExtraTreesRegressor. By combining diverse regression models, VotingRegressor leverages the collective wisdom of individual estimators to improve prediction accuracy and robustness. Although it may not have the highest R^2 and MSE values individually when compared to some of the base estimators, its strength lies in its ability to mitigate the weaknesses of any single model by averaging their predictions, thus potentially enhancing predictive robustness. In our scenario of diesel purity estimation, where accuracy and robustness are crucial, VotingRegressor offers a unified approach that harnesses the strengths of different regression techniques, potentially leading to enhanced predictive performance and more reliable results.

Figure 3 displays the error metrics and accuracy of the VotingRegressor model, demonstrating a close match between simulated and forecasted test data for various percentages of adulterated diesel, with slight fluctuations observed at specific data points. Importantly, the data used for evaluation tests were distinct from those used in the model's training phase.

A more comprehensive statistical analysis of the model is presented in Table 4, where the different models are tested for predicting adulterated diesel concentration. This table depicts the means of error prediction for modified diesel concentration values.

It is noteworthy to highlight that the benchmark dataset labeled in Table 4 has not been employed during either the training or testing phases. The models' average percentage errors are $5.028 \times 10^{-3}\%$, $6.878 \times 10^{-3}\%$, $3.07 \times 10^{-3}\%$, $8.65 \times 10^{-3}\%$, $7 \times 10^{-4}\%$, and $3.87 \times 10^{-3}\%$ for the gradient boosting regressor, decision tree regressor, random forest regressor, linear regression, extra trees regressor, and voting regressor models, respectively.

These findings reveal the models' expanding potential, highlighted by their accuracy in predicting diesel adulteration percentages using data not previously encountered in training.

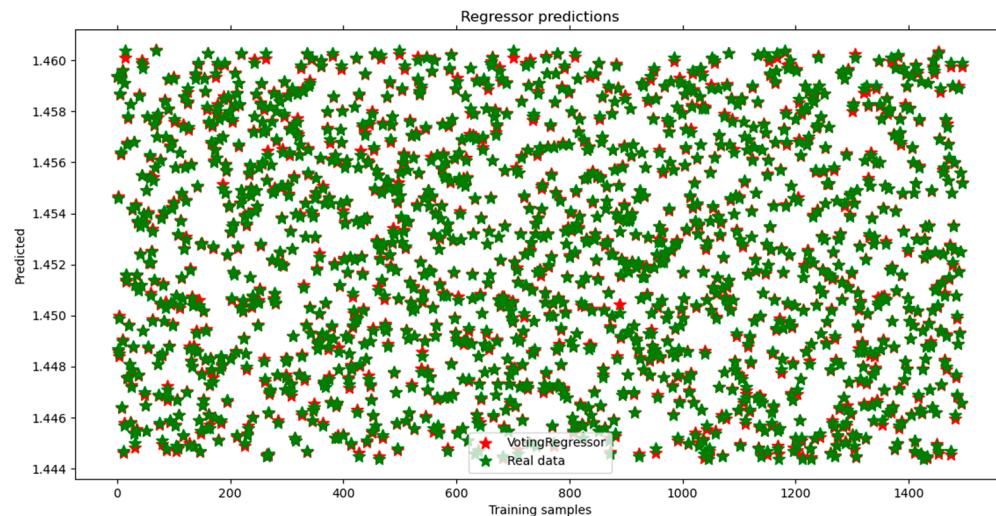


Figure 3. The accuracy of voting regressor model predictions across the 1496 dataset sample test.

Table 4. Analysis of error rates and adulterated diesel percentage predictions for unseen data.

	Real	Predicted	Error %
GradientBoostingRegressor	1.4588	1.45887	0.00476
	1.4491	1.44912	0.0014
	1.4567	1.45678	0.0052
	1.4486	1.4485	0.0069
	1.4538	1.4537	0.00688
DecisionTreeRegressor	1.4588	1.4589	0.00685
	1.4491	1.4492	0.0069
	1.4567	1.4568	0.00686
	1.4486	1.4485	0.0069
	1.4538	1.4537	0.00688
RandomForestRegressor	1.4588	1.45882	0.00171
	1.4491	1.44909	0.00035
	1.4567	1.45671	0.00103
	1.4486	1.44853	0.00511
	1.4538	1.4537	0.00715
LinearRegression	1.4588	1.45926	0.03133
	1.4491	1.44912	0.00116
	1.4567	1.45647	0.01558
	1.4486	1.44892	0.0221
	1.4538	1.45378	0.00127
ExtraTreesRegressor	1.4588	1.45881	0.00082
	1.4491	1.4491	1.07×10^{-13}
	1.4567	1.4567	0.00027
	1.4486	1.44862	0.00138
	1.4538	1.45381	0.00103
VotingRegressor	1.4588	1.45893	0.0091
	1.4491	1.44913	0.00182
	1.4567	1.45665	0.0033
	1.4486	1.44861	0.00091
	1.4538	1.45374	0.00423

Detecting false diesel and measuring relevant concentrations still looks like an under-developed research area, and existing studies show a shortage of emissions data obtained through sensors based on machine learning [19,20] using an optical philosophy. The limited quantity of training data resulted in a decrease in the accuracy of the deep-learning

neural network's ability to correctly classify adulterated diesel, thereby affecting its performance. Our research creates a synthetic dataset that creates the possibility of more accurate forecasts with the help of actual data. As Figure 3 and Table 3 reveal, our model proves that the *d* variable shows a significant correlation with adulterated diesel concentration with additive use labeled, which means we are able to benefit from the generated data without anomalies.

5. Conclusions

This paper outlines an innovative method for detecting adulterated diesel fuel, particularly when mixed with kerosene, by employing refractive index values of both authentic and potentially adulterated diesel samples in conjunction with machine learning algorithms to accurately ascertain the level of adulteration.

- In contrast to traditional detection methods that are expensive, require extensive sample preparation, require skilled technicians, and are not adaptable for field testing or versatile in detecting various diesel adulterants, our suggested approach leverages the principles of optics of light reflection and refraction to create synthetic data for machine learning analysis.
- Our laser-based sensor, designed using COMSOL, is composed of a simple setup involving a diesel-filled container and an overhead laser. The laser light, after refracting through the diesel, is measured for its reflection back to a sensor aligned with the laser.
- Various parameters are calculated, such as the distance from the laser to the point of light detection, angles of incidence, and diesel depth. These parameters are then utilized as synthetic data, streamlining the machine learning training phase, a typically laborious aspect of AI implementation, to predict adulteration levels across a spectrum from 0 to 100%.
- Different models have been tested to check the best performance looking at the hyperparameter metrics. The results validate our models' high accuracy in predicting unseen data, as evidenced by an R-squared value of 0.999 and a mean absolute error of 0.074, confirming their potential for practical application.
- This sensor's cost-effective and versatile design promotes its utility across various applications, making it a promising solution for affordable, low-cost Internet of Things technologies.

The implications of this research are significant, offering advancements in sensor technology for the precise and accessible detection of diesel adulteration. This method is especially advantageous for use in developing countries, where it could significantly diminish the dependence on intricate lab analyses by allowing for initial adulteration screening on-site and thereby reserving detailed lab analyses for only the most challenging samples.

Author Contributions: Conceptualization, B.M. and S.V.; Investigation, B.M. and S.V.; Methodology, B.M. and S.V.; Software, B.M. and S.V.; Supervision, B.M. and S.V.; Validation, B.M. and S.V.; Visualization, B.M., S.V. and T.A.; Writing—original draft, B.M.; Writing—review and editing, B.M., S.V. and T.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

SPR	Surface plasmon resonance
PCF	Photonic crystal fiber
IR	Infrared
PLS	Partial least squares
SVR	Support vector machine regression
MCR-ALS	Multivariate curve resolution with alternating least squares
AI	Artificial intelligence
ML	Machine learning
GEP	Gene expression programming
BSEC	Brake specific energy consumption
NOx	Nitrogen oxides
BTE	Brake thermal efficiency
UHC	Unburned hydrocarbon
CO	Carbon monoxide
NIR	Near-infrared
d	Light path from its point of entry to its intersection with the sensing zone
n_d	Refractive index of diesel
n_k	Refractive index of kerosene
n_{air}	Refractive index of air
n_w	Refractive index of water
θ_i	Incident angle
θ_t	Transmitted angle
θ_r	Reflected angle
W_{air}	Depth of air
W_{diesel}	Depth of diesel sample
λ	Wavelength of light
R^2	R-squared
MSE	Mean squared error
MAE	Mean absolute error
RMSE	Root mean squared error

References

1. Vempatapu, B.P.; Kanaujia, P.K. Monitoring petroleum fuel adulteration: A review of analytical methods. *TrAC—Trends Anal. Chem.* **2017**, *92*, 1–11. [[CrossRef](#)]
2. Mattheou, L.; Zannikos, F.; Schinas, P.; Karavalakis, G.; Karonis, D.; Stournas, S. Impact of using adulterated automotive diesel on the exhaust emissions of a stationary diesel engine. *Glob. NEST J.* **2018**, *8*, 291–296.
3. Nurdin, H.; Hasanuddin, H.; Darmawi, D.; Prasetya, F. Analysis of Calorific Value of Tibarau Cane Briquette. *Mater. Sci. Eng. Conf. Ser.* **2018**, *335*, 012058. [[CrossRef](#)]
4. International Council of Chemical Associations (US/ICCA) COCAM 3. 2012. Available online: <https://hpvchemicals.oecd.org/UI/handler.axd?id=73b56220-3a8b-479b-b03c-99c7353bf4d6> (accessed on 7 April 2024).
5. Yuan, W.; Hansen, A.C.; Zhang, Q. The specific gravity of biodiesel fuels and their blend with diesel fuel. *Agric. Eng. Int. CIGR J. Sci. Res. Dev.* **2004**, *6*. Available online: https://www.researchgate.net/publication/228589856_The_specific_gravity_of_biodiesel_fuels_and_their_blend_with_diesel_fuel (accessed on 7 April 2024).
6. Obuchi, A.; Aoyama, H.; Ohi, A.; Ohuchi, H. Determination of polycyclic aromatic hydrocarbons in diesel exhaust particulate matter and diesel fuel oil. *J. Chromatogr. A* **1984**, *312*, 247–259. [[CrossRef](#)] [[PubMed](#)]
7. Vempatapu, B.P.; Tripathi, D.; Kumar, J.; Kanaujia, P.K. Determination of Kerosene as an Adulterant in Diesel through Chromatography and High-Resolution Mass Spectrometry. *SN Appl. Sci.* **2019**, *1*, 637. [[CrossRef](#)]
8. Chowdhury, M.; Gholizadeh, A.; Agah, M. Rapid Detection of Fuel Adulteration Using Microfabricated Gas Chromatography. *Fuel* **2021**, *286*, 119387. [[CrossRef](#)]
9. Jabin, M.A.; Rana, M.J.; Al-Zahrani, F.A.; Paul, B.K.; Ahmed, K.; Bui, F.M. Novel Detection of Diesel Adulteration Using Silver-Coated Surface Plasmon Resonance Sensor. *Plasmonics* **2022**, *17*, 15–40. [[CrossRef](#)]
10. Moura, H.O.; Câmara, A.B.; Santos, M.C.; Morais, C.L.; de Lima, L.A.; Lima, K.M.; de Carvalho, L.S. Advances in Chemometric Control of Commercial Diesel Adulteration by Kerosene Using IR Spectroscopy. *Anal. Bioanal. Chem.* **2019**, *411*, 2301–2315. [[CrossRef](#)] [[PubMed](#)]

11. Cunha, D.A.; Neto, Á.C.; Colnago, L.A.; Castro, E.V.R.; Barbosa, L.L. Application of Time-Domain NMR as a Methodology to Quantify Adulteration of Diesel Fuel with Soybean Oil and Frying Oil. *Fuel* **2019**, *252*, 149. [[CrossRef](#)]
12. de Aguiar, L.M.; Galvan, D.; Bona, E.; Colnago, L.A.; Killner, M.H.M. Data Fusion of Middle-Resolution NMR Spectroscopy and Low-Field Relaxometry Using the Common Dimensions Analysis (ComDim) to Monitor Diesel Fuel Adulteration. *Talanta* **2022**, *236*, 122838. [[CrossRef](#)] [[PubMed](#)]
13. Cunha, D.A.; Montes, L.F.; Castro, E.V.R.; Barbosa, L.L. NMR in the Time Domain: A New Methodology to Detect Adulteration of Diesel Oil with Kerosene. *Fuel* **2016**, *166*, 78. [[CrossRef](#)]
14. Verma, R.K.; Suwalka, P.; Yadav, J. Detection of Adulteration in Diesel and Petrol by Kerosene Using SPR Based Fiber Optic Technique. *Opt. Fiber Technol.* **2018**, *43*, 11. [[CrossRef](#)]
15. Chauhan, M.; Khanikar, T.; Singh, V.K. PDMS Coated Fiber Optic Sensor for Efficient Detection of Fuel Adulteration. *Appl. Phys. B* **2022**, *128*, 109. [[CrossRef](#)]
16. Roy, S. Fiber Optic Sensor for Determining Adulteration of Petrol and Diesel by Kerosene. *Sens. Actuators B Chem.* **1999**, *55*, 171. [[CrossRef](#)]
17. Bell, J.; Gotor, R.; Rurack, K. Fluorescent Paper Strips for the Detection of Diesel Adulteration with Smartphone Read-Out. *J. Vis. Exp.* **2018**, *141*, 58019.
18. Kanyathare, B.; Kuivalainen, K.; Räty, J.; Silfsten, P.; Bawuah, P.; Peiponen, K.E. A Prototype of an Optical Sensor for the Identification of Diesel Oil Adulterated by Kerosene. *J. Eur. Opt. Soc.* **2018**, *14*, 71. [[CrossRef](#)]
19. Sadat, A. Determining the Adulteration of Diesel by an Optical Method. *Int. J. Comput. Appl.* **2014**, *100*, 17588. [[CrossRef](#)]
20. Paiva, E.M.; Rohwedder, J.J.R.; Pasquini, C.; Pimentel, M.F.; Pereira, C.F. Quantification of Biodiesel and Adulteration with Vegetable Oils in Diesel/Biodiesel Blends Using Portable Near-Infrared Spectrometer. *Fuel* **2015**, *160*, 67. [[CrossRef](#)]
21. Barra, I.; Mansouri, M.A.; Bousrabat, M.; Cherrah, Y.; Bouklouze, A.; Kharbach, M. Discrimination and Quantification of Moroccan Gasoline Adulteration with Diesel Using Fourier Transform Infrared Spectroscopy and Chemometric Tools. *J. AOAC Int.* **2019**, *102*, 966–970. [[CrossRef](#)] [[PubMed](#)]
22. Pontes, M.J.C.; Pereira, C.F.; Pimentel, M.F.; Vasconcelos, F.V.C.; Silva, A.G.B. Screening Analysis to Detect Adulteration in Diesel/Biodiesel Blends Using Near Infrared Spectrometry and Multivariate Classification. *Talanta* **2011**, *85*, 2159–2165. [[CrossRef](#)]
23. Kanyathare, B.; Asamoah, B.; Peiponen, K.E. Imaginary Optical Constants in Near-Infrared (NIR) Spectral Range for the Separation and Discrimination of Adulterated Diesel Oil Binary Mixtures. *Opt. Rev.* **2018**, *26*, 85–94. [[CrossRef](#)]
24. Kumar, A.; Singh, V.R.; Parashar, D.C. Ultrasonic Detection of Adulteration in Diesel. *Res. Ind.* **1991**, *36*, 168–170.
25. Bhowmik, S.; Paul, A.; Panua, R.; Ghosh, S.K.; Debroy, D. Artificial Intelligence Based Gene Expression Programming (GEP) Model Prediction of Diesel Engine Performances and Exhaust Emissions Under Diesosenol Fuel Strategies. *Fuel* **2019**, *235*, 317–325. [[CrossRef](#)]
26. Babu, V.; Krishna, R.; Mani, N. Review on the Detection of Adulteration in Fuels through Computational Techniques. *Mater. Today Proc.* **2017**, *4*, 1723–1729. [[CrossRef](#)]
27. De Matos, T.S.; Dos Santos, R.C.; De Souza, C.G.; De Carvalho, R.C.; De Andrade, D.F.; D’ávila, L.A. Determination of the Biodiesel Content on Biodiesel/Diesel Blends and Their Adulteration with Vegetable Oil by High-Performance Liquid Chromatography. *Energy Fuels* **2019**, *33*, 11310–11317. [[CrossRef](#)]
28. Ejilah, I.R.; Olorunnishola, A.A.G.; Enyejo, L.A. A Comparative Analysis of the Combustion Behavior of Adulterated Kerosene Fuel Samples in a Pressurized Cooking Stove. *Glob. J. Res. Eng. Mech. Eng.* **2013**, *13*, 34–44.
29. de Vasconcelos, F.V.C.; de Souza, P.F.B.; Pimentel, M.F.; Pontes, M.J.C.; Pereira, C.F. Using Near-Infrared Overtone Regions to Determine Biodiesel Content and Adulteration of Diesel/Biodiesel Blends with Vegetable Oils. *Anal. Chim. Acta* **2012**, *716*, 101–107. [[CrossRef](#)] [[PubMed](#)]
30. Ogundare, F.; Adekola, F.; Oladosu, I. Compositions and photon mass attenuation coefficients of diesel, kerosene, palm and groundnut oils. *Fuel* **2019**, *255*, 115697. [[CrossRef](#)]
31. Tran, N.; Chen, H.; Bhuyan, J.; Ding, J. Data Curation and Quality Evaluation for Machine Learning-Based Cyber Intrusion Detection. *IEEE Access* **2022**, *10*, 121900–121923. [[CrossRef](#)]
32. Mourched, B.; Abdallah, M.; Hoxha, M.; Vrtagic, S. Machine-Learning-Based Sensor Design for Water Salinity Prediction: A Conceptual Approach. *Sustainability* **2023**, *15*, 11468. [[CrossRef](#)]
33. Demircioğlu, U.; Sayıl, A.; Bakır, H. Detecting Cutout Shape and Predicting Its Location in Sandwich Structures Using Free Vibration Analysis and Tuned Machine-Learning Algorithms. *Arab. J. Sci. Eng.* **2023**, *49*, 1611–1624. [[CrossRef](#)]
34. Chugh, S.; Ghosh, S.; Gulistan, A.; Rahman, B.M.A. Machine Learning Regression Approach to the Nanophotonic Waveguide Analyses. *J. Light. Technol.* **2019**, *37*, 6080–6089. [[CrossRef](#)]
35. Mourched, B.; Hoxha, M.; Abdelgalil, A.; Ferko, N.; Abdallah, M.; Potams, A.; Lushi, A.; Turan, H.I.; Vrtagic, S. Piezoelectric-Based Sensor Concept and Design with Machine Learning-Enabled Using COMSOL Multiphysics. *Appl. Sci.* **2022**, *12*, 9798. [[CrossRef](#)]
36. Wang, Y.; Guo, J.; Yang, Z.; Dou, Y.; Chang, X.; Sun, R.; Zuo, G.; Yang, W.; Liang, C.; Hao, Y.; et al. Computer Prediction of Seawater Sensor Parameters in the Central Arctic Region Based on Hybrid Machine Learning Algorithms. *IEEE Access* **2020**, *8*, 213783–213798. [[CrossRef](#)]
37. Ray Optics Module User’s Guide. COMSOL Multiphysics®v. 6.2. COMSOL AB, Stockholm, Sweden. 2023. Available online: <https://doc.comsol.com/5.4/doc/com.comsol.help.roptics/RayOpticsModuleUsersGuide.pdf> (accessed on 7 April 2024).

38. Bhausaheb, M. Determination of Adulteration in Diesel by Refractive Index Measurements. *Int. J. Appl. Chem.* **2008**, *4*, 247–252.
39. Kanyathare, B.; Peiponen, K.E. Hand-Held Refractometer-Based Measurement and Excess Permittivity Analysis Method for Detection of Diesel Oils Adulterated by Kerosene in Field Conditions. *Sensors* **2018**, *18*, 1551. [CrossRef] [PubMed]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.