

Article

A Comparative Study of Machine Learning Algorithms for Measuring Thin-Film Thickness Using Terahertz Time-Domain Waves Simulated by the Finite Difference Time Domain Method

Pingan Liu ¹, Xiangjun Li ^{2,*}, Yibing Liu ¹ and Liguao Zhu ^{1,*}

¹ Institute of Fluid Physics, China Academy of Engineering Physics, Mianyang 621900, China; liupingan23@gscaep.ac.cn (P.L.); liuyibing20@gscaep.ac.cn (Y.L.)

² Centre for THz Research, China Jiliang University, Hangzhou 310018, China

* Correspondence: xiangjun_li@cjl.u.edu.cn (X.L.); zhuliguao@tsinghua.org.cn (L.Z.)

Abstract

Terahertz (THz) waves offer unique advantages, including non-contact operation, high penetration capability, and high resolution, making them particularly well-suited for the non-destructive thickness measurement of film-structured materials. In reflective terahertz time-domain spectroscopy (THz-TDS), thickness measurement approaches are generally classified into three categories: optimization-based methods that rely on theoretical models, time-of-flight (ToF), and machine learning. Model-based optimization techniques require precise knowledge of the optical parameters and structural configuration of each layer; however, they often suffer from slow convergence and are prone to becoming trapped in local optima. In contrast, ToF-based methods determine thickness by calculating the time delay between echo pulses reflected from different interfaces, yet their applicability is limited when the film thickness is extremely small. Machine learning, especially deep learning, enables the establishment of a direct, data-driven mapping between THz waveforms (or their extracted features) and the target thickness. Such approaches offer rapid inference, strong robustness to noise, and good adaptability to thin or structurally complex films, although their accuracy remains dependent on the quality of training data and the generalization capability of the model. In this study, high-fidelity THz waveform data generated via finite-difference time-domain (FDTD) simulations are utilized to conduct a comparative investigation into the film thickness prediction performance of several representative machine learning algorithms, including Back Propagation (BP) neural networks, Support Vector Machines (SVM), Random Forests (RF), Extreme Learning Machines (ELM), K-Nearest Neighbors (KNN), and Partial Least Squares (PLS) regression. The results indicate that, in terms of prediction error, the overall ranking of algorithmic performance from best to worst is: PLS > RF > SVM > BP > ELM > KNN. These findings provide valuable guidance for the future application of machine learning-assisted THz-TDS in precise film thickness measurement.



Academic Editor: Paolo Mele

Received: 6 July 2026

Revised: 30 July 2026

Accepted: 3 August 2026

Published: 4 August 2026

Copyright: © 2026 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: Terahertz Time-Domain Spectroscopy (THz-TDS); Finite Difference Time Domain (FDTD); thin-film thickness measurement; machine learning; PARTIAL Least Squares (PLS); non-destructive testing; Discrete Cosine Transform (DCT)

1. Introduction

The precise measurement of thin film thickness is critical across numerous industrial sectors, including aerospace thermal barrier coating inspection, semiconductor process

monitoring, automotive coating quality control, and lubricating oil film evaluation [1]. Achieving non-contact, high-precision, and rapid thickness characterization remains a key technological challenge for broad industrial adoption. Terahertz electromagnetic waves have demonstrated strong penetration capabilities through a wide range of dielectric materials. Compared with frequency-domain spectroscopy, THz-TDS offers significant advantages for thickness measurement, including broader measurable ranges and the ability to characterize complex multilayer structures [2,3].

Existing thickness retrieval methods based on THz-TDS can generally be classified into three categories: model inversion, time-of-flight analysis, and machine learning approaches. Model inversion relies heavily on the accuracy of both the physical model and the optical parameters of each layer, often suffering from slow convergence and susceptibility to local minima [4,5]. Time-of-flight methods determine thickness from the temporal separation between reflected pulses at different interfaces but are inherently limited in resolving ultra-thin films [6]. In contrast, machine learning—particularly deep learning—exploits complex mappings between THz waveform features to enable fast, robust, and accurate predictions, even for thin or structurally complex films. Nevertheless, when films are extremely thin ($<10\text{ }\mu\text{m}$) or consist of multiple layers with poorly defined interfaces, conventional model inversion and time-of-flight algorithms exhibit considerable instability, limiting the practical deployment of terahertz techniques for ultra-thin film metrology [2,7].

Considerable research efforts have been devoted to applying machine learning to THz-TDS-based film measurement. For instance, Xu et al. proposed a hybrid Principal Component Analysis (PCA)–Particle Swarm Optimization (PSO)–ELM model that achieved outstanding predictive performance, with an R^2 value close to 1 and minimal prediction error [8]; they further developed an attention-based Long Short-Term Memory (LSTM)–Support Vector Regression fusion model, reducing the root-mean-square prediction error to $8.9\text{ }\mu\text{m}$ [9]. Li et al. implemented an optimized Elman neural network within a hybrid framework, attaining an R^2 of 0.999 [10]. Sun et al. introduced a model-data-driven interpretable convolutional neural network, achieving a mean absolute error of $1.73 \pm 0.13\text{ }\mu\text{m}$ [11]. Jiang et al. demonstrated that a trained BP neural network could predict coating thickness with over 96% accuracy [12]. Kaviya et al. applied a PCA–Whale Optimization Algorithm (WOA)–Elman hybrid model, reporting an R^2 of 0.999 for thickness prediction [13]. Wu et al. integrated physics-informed neural networks with the Rouard model, pushing the minimum measurable thickness to $4.5\text{ }\mu\text{m}$ [14].

In summary, these studies highlight the flexibility and applicability of machine learning—especially deep learning—for THz-TDS film metrology. However, most existing works focus on individual algorithms or narrow model families, with limited comparative analysis among multiple approaches. Moreover, available datasets are predominantly derived from experimental measurements, which are often constrained in quantity and diversity. To address these limitations, this study evaluates multiple machine learning and deep learning algorithms [15]—including BP, SVM, RF, ELM, KNN, and PLS. Synthetic THz-TDS waveforms were generated via finite-difference time-domain (FDTD) simulations of single-layer films on dielectric substrates, using normally incident terahertz plane waves. To emulate realistic measurement conditions, Gaussian white noise was added to the simulated time-domain signals, and the discrete cosine transform (DCT) was employed for dimensionality reduction. Predictive performance was assessed using large-scale simulated datasets that substantially exceed typical experimental sample size. Results indicate that, ranked by prediction error from best to worst, the algorithms perform as follows: PLS > RF > SVM > BP > ELM > KNN. These findings provide a robust foundation for the informed selection and optimization of machine learning approaches for thin-film measurement based on THz-TDS.

2. Structure and Methods

As illustrated in Figure 1a, the thin-film thickness is assumed to be measured using a THz-TDS system operating in reflection mode. This system employs a femtosecond fiber laser centered at 1560 nm, delivering pulses shorter than 100 fs at a repetition rate of 100 MHz. Terahertz radiation is generated and coherently detected via photoconductive antennas. A fiber-optical delay line provides a scan range of up to 300 ps at a scanning speed of 20 Hz. The system operates over an effective frequency range of 0.1–3.0 THz, with a dynamic signal-to-noise ratio exceeding 60 dB.

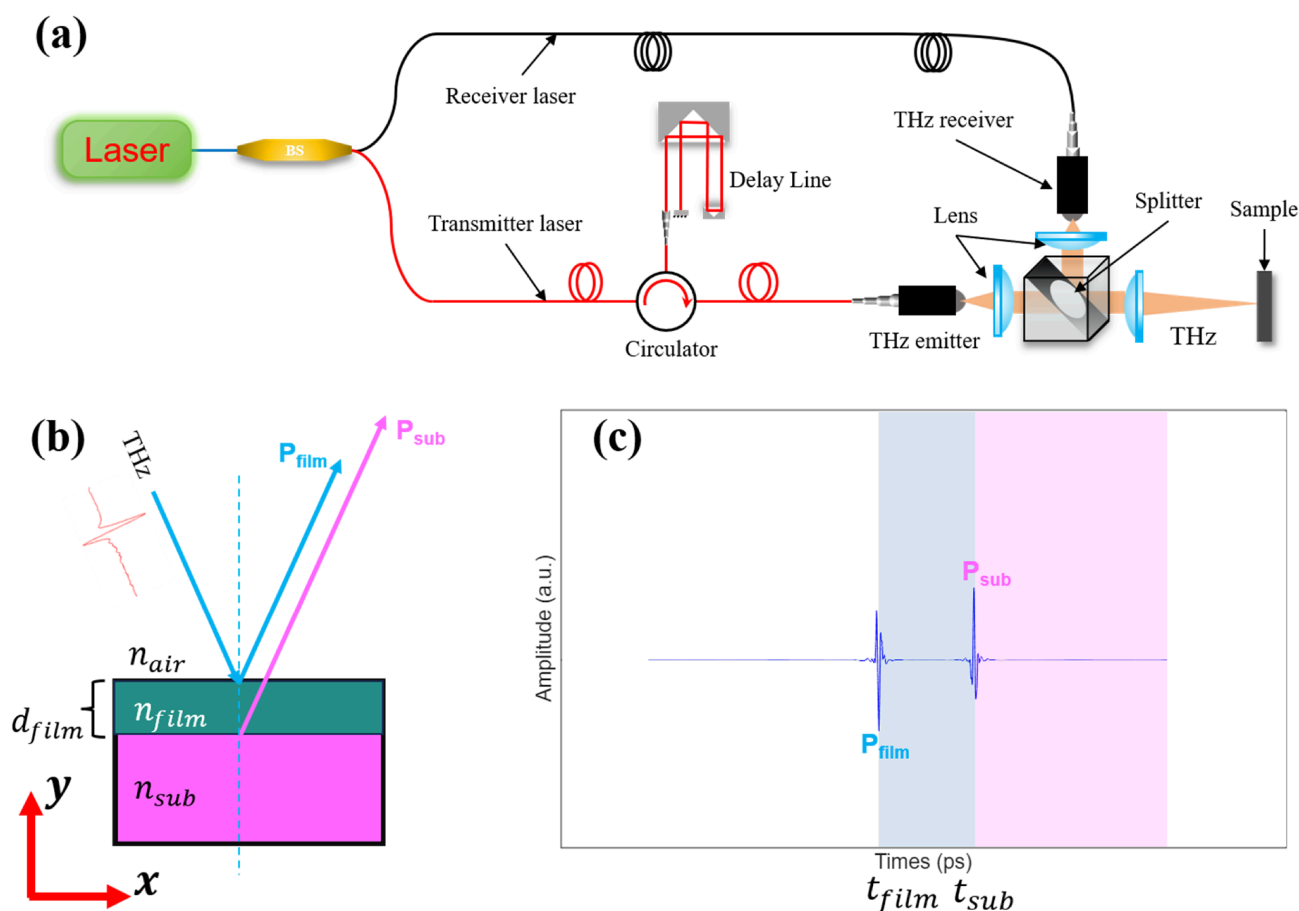


Figure 1. The film measurement model based on a THz-TDS reflection setup: (a) THz-TDS reflection setup; (b) diagram of one-layer film on the substrate; (c) the reflected THz-TDS pulses of the film structure.

The measurement target consists of a single-layer film deposited on a dielectric substrate, forming a two-layer structure illustrated in Figure 1b. In this study, the film is modeled as a plastic material with a refractive index $n = 1.5$ and an extinction coefficient $k = 0.03$. At the same time, the substrate is represented as a composite material with a refractive index $n = 3.5$ and an extinction coefficient $k = 0.05$. In reflection mode, a vertically incident picosecond THz pulse interacts with the sample, producing a reflected signal composed primarily of two contributions: the first reflection from the air–film interface (P_{film}) and the second reflection from the film–substrate interface (P_{sub}), as depicted in Figure 1c. The conventional time-of-flight analysis relies on accurately resolving the time interval between t_{film} and t_{sub} to calculate the film thickness from its refractive index. However, for ultra-thin films, the two reflected pulses overlap significantly in the time domain, rendering traditional approaches unstable and unreliable. At the same time, the model inversion method will face the non-convergent problem.

To address this limitation, a machine learning-based thickness prediction framework with multiple machine learning (ML) methods is introduced, as outlined in the workflow diagram in Figure 2. The procedure begins with FDTD simulations of the THz reflection response from the film–substrate structure under normal incidence. The incident time-domain waveform experimentally generated by the THz-TDS system in Figure 1a and its Fourier transform are shown in Figure 3a,b. In the FDTD simulation, the waveform is set as a plane-wave source vertically incident on the film structure. The periodic boundaries are applied on the x-axis. The perfectly matched layer boundaries are applied on the y-axis. When the film thickness is assumed in the range of 20~100 μm in steps of 20 μm , the computed reflection waveforms are demonstrated in Figure 3c.

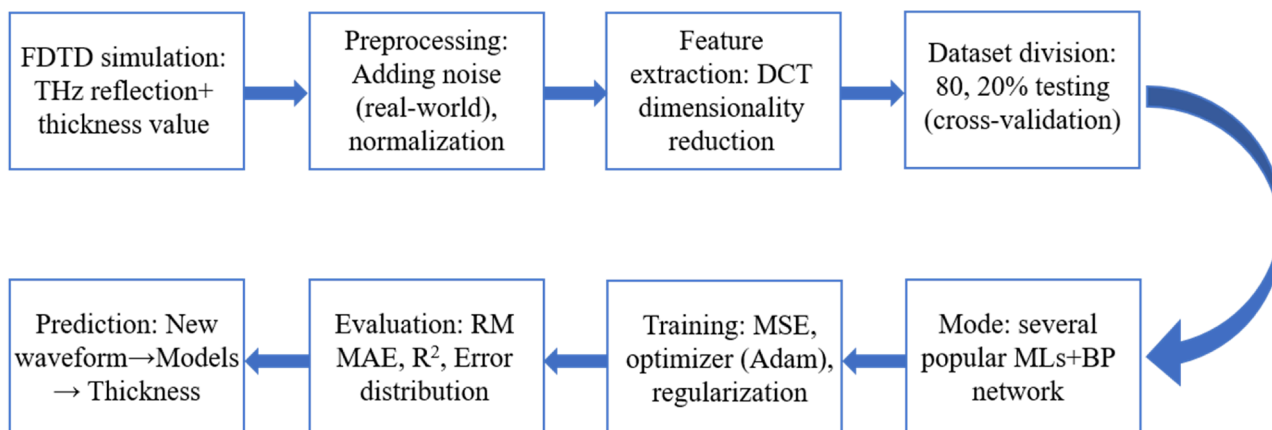


Figure 2. The workflow diagram of predicting thin-film thickness using machine learning algorithms by THz-TDS simulated data.

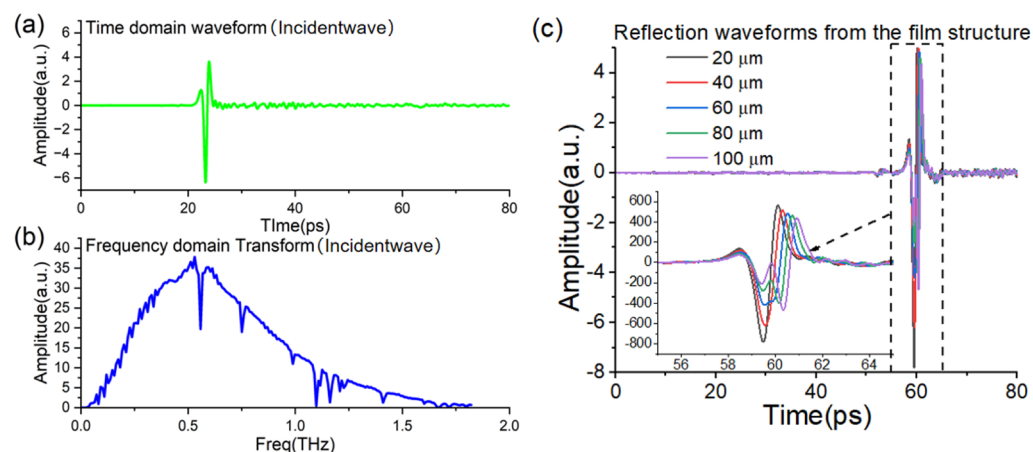


Figure 3. The exciting and simulated waveforms in the FDTD simulation. (a) The exciting waveform; (b) the frequency domain transform; (c) computed reflection waveforms of the film in 20~100 μm thickness in steps of 20 μm .

For the training and validation of the machine learning models, 2000 simulated reflection signals were generated using FDTD by randomly varying the film thickness between 0 and 200 μm , as depicted in Figure 4a. To emulate realistic measurement conditions, the signal dataset was augmented with 20 dB Gaussian white noise and subsequently normalized to ensure consistent scaling. As shown in Figure 4b, the black solid line represents the original simulated reflection waveform, while the pink solid line corresponds to the signal contaminated with 20 dB Gaussian white noise. The Discrete Cosine Transform (DCT) was then applied to the noisy time-domain signals to reduce dimensionality while

retaining the most informative spectral features. Figure 4c,d illustrate the first 1000 DCT coefficients derived from the original signal with and without added noise, respectively. The results indicate that the primary energy of the signals is concentrated within the 1–400 coefficient range. Furthermore, as demonstrated in the inset of Figure 4b, the reconstructed waveform (blue dashed line) utilizing only 400 DCT components exhibits a high degree of consistency with the original signal. These findings confirm that the DCT effectively reduces data volume while mitigating the impact of noise.

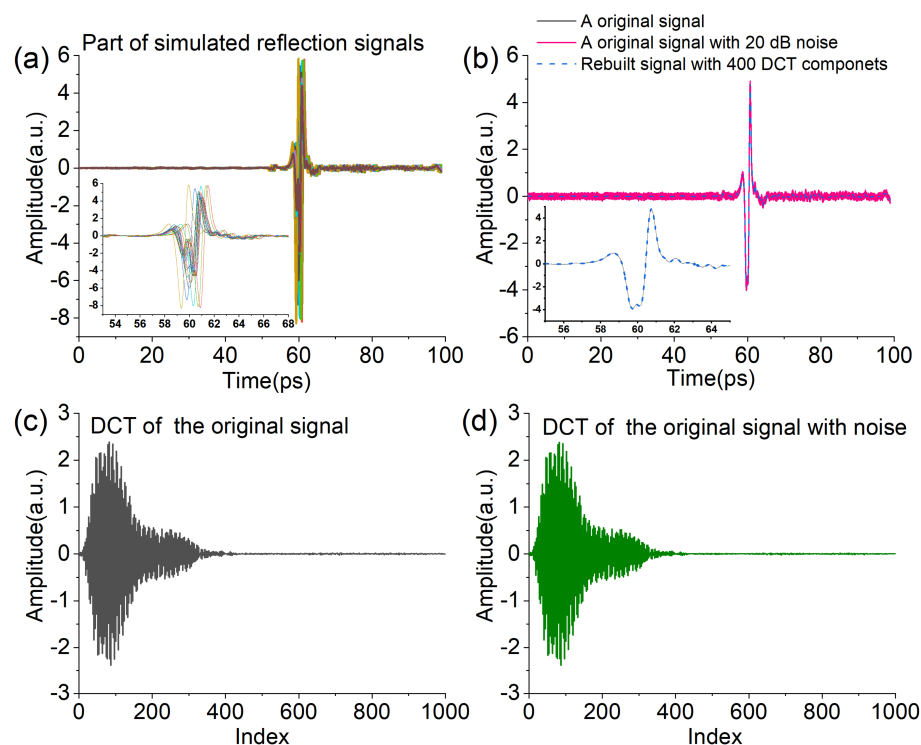


Figure 4. The simulated reflection waveforms by the film structure and the DCT transformation. (a) Part of simulated reflection signals; (b) one simulated signal adding noise; (c) DCT of the original signal; (d) DCT of the original signal with noise.

Then, the resulting dataset is partitioned into training (80%) and testing (20%) subsets, with cross-validation applied during model development. Multiple regression models are evaluated, including BP, SVM, RF, ELM, KNN, and PLS methods. Model training employs the mean squared error (MSE) as the loss function, with optimization performed using the proper optimizer and regularization strategies to mitigate overfitting. Trained models are assessed using standard performance metrics—root-mean-square error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), and prediction error distributions—to identify the most accurate and robust approach. Once validated, the optimal model is deployed to predict the thickness of unknown samples directly from their THz reflection waveforms. By integrating physics-based simulation, signal processing, and machine learning, this methodology overcomes the resolution limits of conventional THz-TDS thickness measurement and enables accurate characterization of ultra-thin films in a non-contact, rapid, and scalable manner.

3. Results and Discussion

Predicting film thickness from THz-TDS waveforms constitutes a multivariate regression problem that can be addressed using supervised learning algorithms. Six representative methods were selected for comparison: Backpropagation Neural Networks (BP), Support Vector Regression (SVR), Random Forest (RF), Extreme Learning Machines (ELM),

K-Nearest Neighbors (KNN), and Partial Least Squares (PLS). These algorithms collectively cover the major paradigms of mainstream machine learning, and their characteristics are summarized below.

BP Networks comprise multiple layers of interconnected neurons with nonlinear activation functions, enabling universal approximation of continuous functions. Their high flexibility allows them to capture complex patterns in large, high-dimensional datasets. The primary challenges are substantial data requirements and high computational costs during training. Additionally, BP networks suffer from low interpretability, often operating as complete “black boxes.”

SVR aims to fit a function within an ϵ -insensitive tube, penalizing only deviations outside this margin. Through kernel functions, SVR effectively models nonlinear relationships even in high-dimensional feature spaces. Its strengths include strong theoretical grounding and robust generalization performance under small sample conditions. However, SVR scales poorly to large datasets, demands meticulous tuning of kernel parameters, and exhibits high sensitivity to feature scaling, limiting its practicality in large-scale applications.

RF constructs an ensemble of decorrelated decision trees via bootstrap sampling and random feature selection, aggregating their outputs to improve robustness and mitigate overfitting. It also yields intrinsic measures of feature importance. The main drawbacks are reduced transparency due to the ensemble nature and increased computational overhead during training and inference, although it may still encounter difficulties with extremely noisy data.

ELM is a single-hidden-layer feedforward neural network where input weights and hidden biases are randomly assigned and fixed, while output weights are analytically determined in a single step via the Moore–Penrose generalized inverse. This eliminates iterative tuning, yielding extremely fast training speeds, simple implementation, and minimal hyperparameter optimization, while maintaining good generalization. Nevertheless, reliance on random initialization introduces instability and sensitivity to the number of hidden neurons, accompanied by limited interpretability, weak robustness to noise, and a lack of native support for deep or sequential learning.

KNN Regression is a non-parametric method that predicts target values based on the average of the nearest training instances using a specified distance metric. Its principal advantage is flexibility, imposing no functional form and approximating arbitrary nonlinear relationships. However, this simplicity incurs high computational cost during prediction, particularly for large datasets, as distances must be recomputed for each query. KNN is also highly sensitive to feature scaling, noise, and the curse of dimensionality, necessitating rigorous preprocessing.

PLS Regression projects both predictors and responses into a latent space to maximize their covariance. By prioritizing directions most relevant to the prediction target, PLS effectively addresses multicollinearity and high dimensionality, outperforming ordinary regression when the number of predictors exceeds the number of observations—a common scenario in chemometrics and bioinformatics. Limitations include reduced interpretability relative to standard linear regression and sensitivity to outliers, which can distort the extracted components.

After thorough computation and comparison, the optimized parameters for the ML methods BP, SVM, RF, ELM, KNN, and PLS are listed in Table 1. We optimized the hyperparameters for each method via grid search over broad predefined ranges. Model performance was evaluated using 5-fold cross-validation, with results averaged across five runs. The training and prediction (inference) were run in MATLAB 2025b with a fixed random seed for reproducibility on a computer with an AMD Ryzen 9 9950X3D 16-Core Processor (4.30 GHz) and 64GB DDR5 RAM. Figure 5. presents a comprehensive

evaluation of the BP model's prediction performance using four distinct visualization methods. Figure 5a compares the true thickness values (solid red line) with the predicted values (dashed blue line) sequentially by sample index, illustrating the model's stability and consistency over successive predictions. Furthermore, in Figure 5b the horizontal axis represents the ground truth values, while the vertical axis indicates the predictions made by BP network. The tightly clustered data points along the diagonal line indicate a strong linear correlation between the predicted and actual values, suggesting high precision and reliability of the model across the entire measurement range (0–100 μm). Figure 5c analyzes the error distribution by plotting the residuals (the difference between true and predicted values) against the predicted values. The red dashed line at zero represents the ideal scenario with no error. The random scattering of data points around these zero lines, without any discernible trend or systematic bias, confirms that the model does not exhibit overfitting or significant heteroscedasticity across different value ranges. Similarly, the histogram in Figure 5d provides a statistical summary of the prediction errors. The distribution is approximately Gaussian normal and is sharply centered at 0 μm , with the vast majority of residuals falling within the ± 0.5 μm range. This bell-shaped curve signifies that the prediction errors are minimal and randomly distributed, which is a key indicator of a well-generalized regression model.

Table 1. The optimized parameters for ML methods of BP, SVM, RF, ELM, KNN, and PLS.

Algorithm	Key Hyperparameter	Value
BP	Hidden Layer Size/Neurons	5 layers (512/256/128/64/32)
	Activation Function	ReLU
	Solver/Optimizer	Adam
	Learning Rate	0.001
	Max Iterations	100
SVR	Kernel	Linear
	C (Regularization)	10
	Gamma (Kernel Coef.)	0.001
	Epsilon (ϵ -tube)	0.1
RF	n_estimators	500
	max_depth	None
	min_samples_split	2
ELM	n_hidden	1000
	Activation Function	ReLU
	Regularization (C)	0.1
KNN	n_neighbors (k)	1
	Weights	uniform
	Metric	euclidean
PLS	n_components	500
	Scale	TRUE

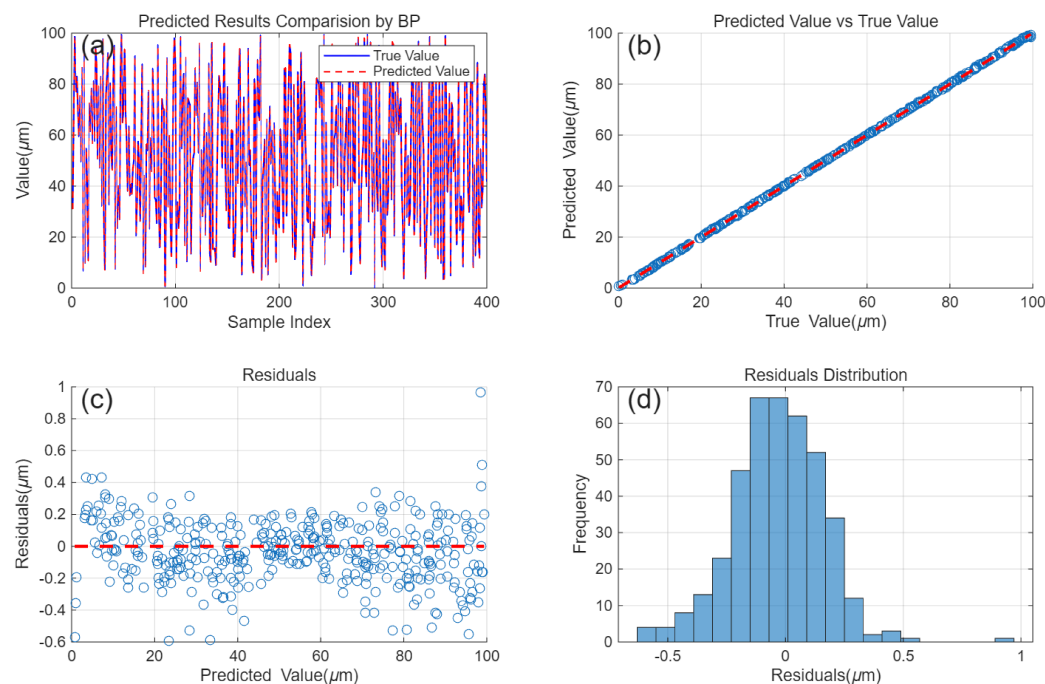


Figure 5. The predicting performance of BP model (a) Predicted results comparison; (b) predicted value vs. true value; (c) residuals vs. predicted value; (d) residuals distribution.

An overall comparison of the six regression models (PLS, RF, BP, SVM, ELM, and KNN) based on their residual distributions in prediction accuracy, stability, and generalization capability for film-thickness inversion is displayed in Figures 6 and 7. Obviously, PLS delivers the most compelling performance, with residuals tightly compressed within an exceptionally narrow band (~ -0.12 to $+0.12$ μm) and no discernible systematic drift across the full 0–100 μm range. Its superior behavior is purpose-built for high-collinearity spectral data, projecting the informative variance into latent variables that simultaneously suppress noise and avoid overfitting. Closely behind, RF achieves the second-best profile, confining most residuals within $\sim \pm 0.2$ μm and maintaining a dense, symmetric scatter around zero. The absence of structural patterning in its residuals signals robust generalization; the slight widening relative to PLS is an expected trade-off of tree-based ensemble variance, but it remains well-separated from the bottom half of the field. The middle tier is anchored by BP and SVM. Both exhibit genuinely random residual scatter, but they begin to show tell-tale signs of higher variance toward ± 0.5 μm , and a visibly broader dispersion envelope than PLS/RF. BP's nonlinear activation structure possesses more expressive capacity to track subtle waveform–thickness nonlinearities, whereas SVM's spread suggests the chosen kernel/regularization combination is slightly under-smoothing or over-penalizing edge cases. Deeper into the lower tier, ELM produces a diffusely scattered residual cloud with consistently larger swings than BP/SVM, reflecting a known ELM weakness. Random hidden-node initialization yields low training cost but can leave the solution sitting in a flatter, noisier region of the loss landscape, increasing prediction variance when the input dimension is already compressed. Finally, KNN are the most widely scattered, with frequent extreme deviations (up to $\sim \pm 1.5$ μm) and no meaningful concentration near zero. This is characteristic of KNN's fundamental limitation with no intrinsic denoising or spectral abstraction mechanism, so the 20 dB Gaussian noise injected into the time-domain signals propagates directly into unstable nearest-neighbor matches after DCT projection.

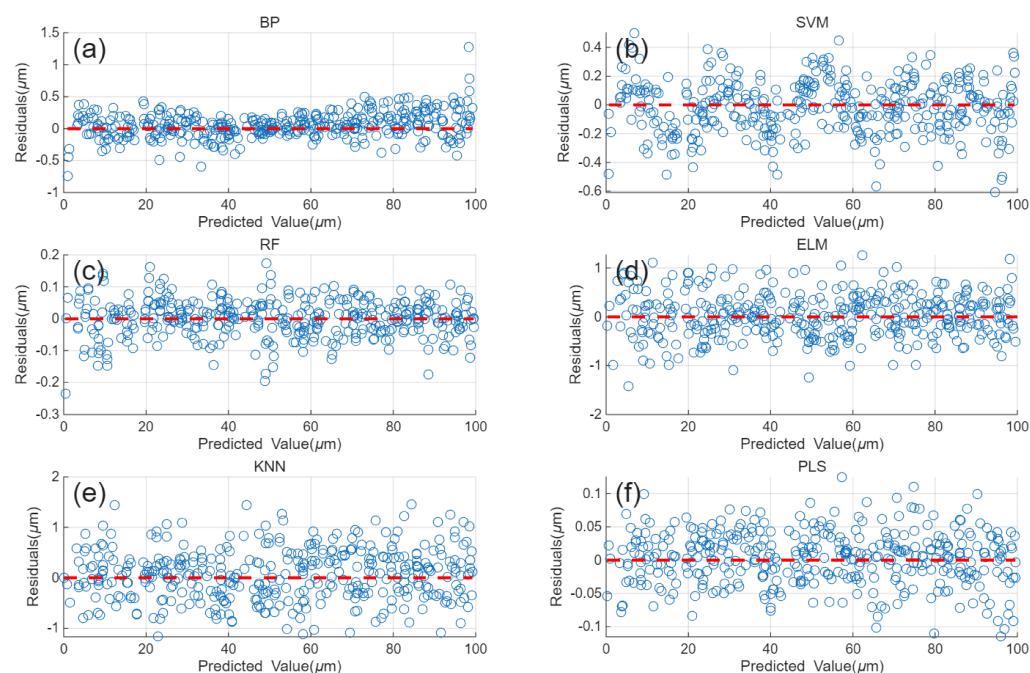


Figure 6. Residuals vs. predicted values for the selected ML methods. (a) BP; (b) SVM; (c) RF; (d) ELM; (e) KNN; (f) PLS.

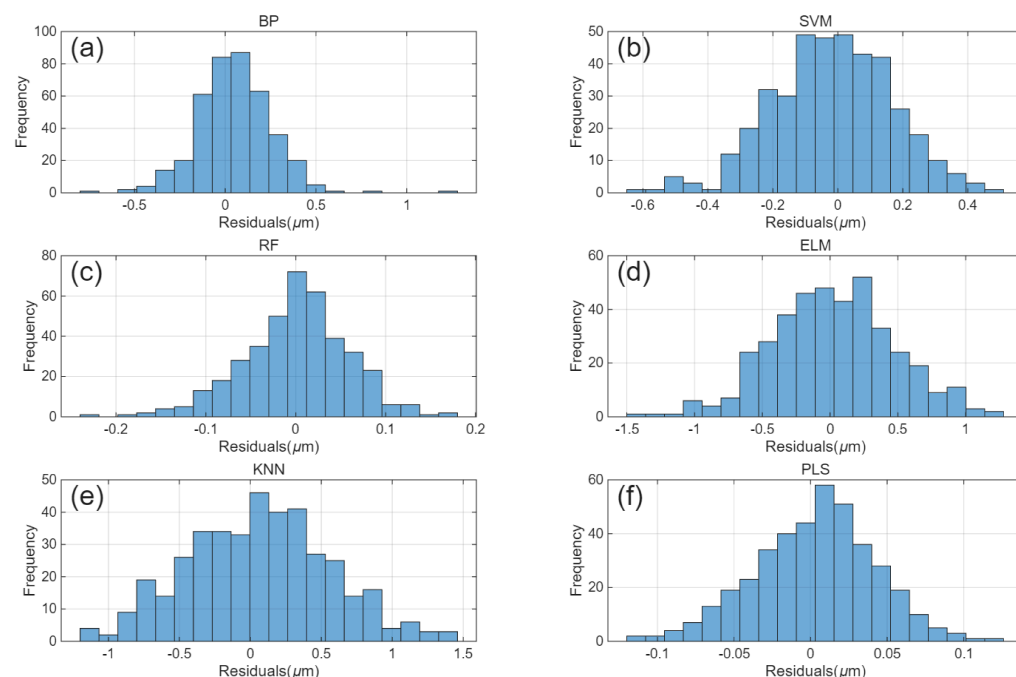


Figure 7. Residual distributions of the selected ML methods. (a) BP; (b) SVM; (c) RF; (d) ELM; (e) KNN; (f) PLS.

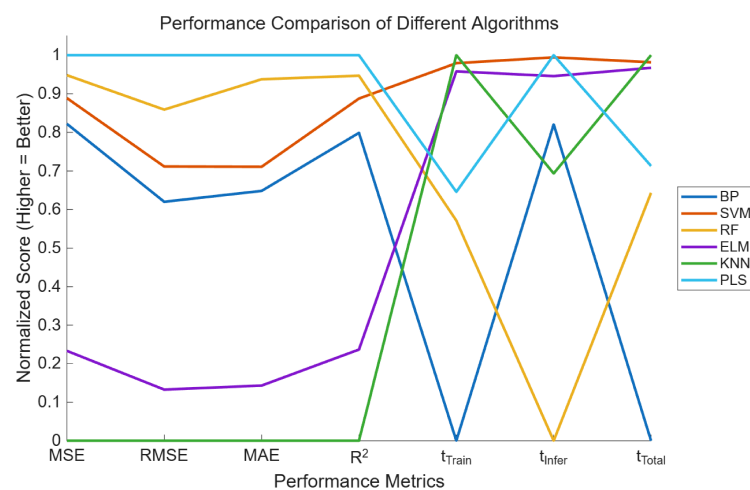
Furthermore, the performance indicators of the 6 ML algorithms with precision parameters such as MSE, RMSE, MAE, R^2 and efficiency parameters such as training time (t_{train}), inferring time (t_{infer}), and total time of both training and inferring (t_{total}) are listed in Table 2. All models achieve a high coefficient of determination with R^2 greater than 0.9996, indicating an excellent overall fit to the data. However, examining the error metrics reveals clear distinctions. According to precision, PLS and RF are the top performers. PLS achieves the lowest error values across the board, making it the most accurate model. RF follows closely with PLS. Intermediate SVM and BP demonstrate moderate accuracy.

Table 2. The performance metrics of the selected ML methods.

	BP	SVM	RF	ELM	KNN	PLS
MSE	0.050 ± 0.008	0.032 ± 0.004	0.016 ± 0.024	0.209 ± 0.015	0.272 ± 0.016	$0.0021 \pm 2 \times 10^{-4}$
RMSE	0.222 ± 0.019	0.178 ± 0.010	0.107 ± 0.076	0.457 ± 0.016	0.521 ± 0.015	0.039 ± 0.002
MAE	0.165 ± 0.017	0.141 ± 0.008	0.054 ± 0.010	0.359 ± 0.015	0.414 ± 0.014	0.030 ± 0.002
R ²	$0.99993 \pm 1 \times 10^{-5}$	$0.99996 \pm 5 \times 10^{-6}$	$0.99998 \pm 3 \times 10^{-5}$	$0.99974 \pm 3 \times 10^{-5}$	$0.99966 \pm 3 \times 10^{-5}$	$0.99999 \pm 2 \times 10^{-7}$
t _{train} (s)	8.125 ± 0.253	0.201 ± 0.018	3.509 ± 0.010	0.373 ± 0.006	0.032 ± 0.006	2.906 ± 0.018
t _{infer} (s)	0.016 ± 0.009	$0.0011 \pm 1 \times 10^{-4}$	0.087 ± 0.008	$0.0052 \pm 2 \times 10^{-4}$	0.027 ± 0.005	$5 \times 10^{-4} \pm 3 \times 10^{-5}$
t _{total} (s)	9.895 ± 3.012	0.237 ± 0.039	3.563 ± 0.019	0.378 ± 0.007	0.055 ± 0.002	2.879 ± 0.017

ELM and KNN exhibit the poorest fitting precision. Both models register MSE values above 0.2, which are an order of magnitude higher than those of PLS, indicating they struggle to capture the underlying data patterns effectively. On the other hand, considering the computational efficiency, PLS is the fastest algorithm by a significant margin, requiring only $\sim 5 \times 10^{-4}$ s for prediction. ELM and KNN are also very fast, while RF is the slowest during inference, likely due to the overhead of traversing multiple decision trees. KNN is the fastest to train as it involves no actual learning process, merely storing data. SVM is also relatively efficient. Conversely, BP is the most time-consuming to train, reflecting the computational intensity of iterative backpropagation, followed by RF.

The comparative results in Figure 8 demonstrate a significant divergence in performance among the evaluated regression models, highlighting the trade-offs between physical interpretability, numerical precision, and computational efficiency in THz-TDS inversion. Beyond the empirical performance comparison, the consistent superiority of PLS regression for THz-TDS data can be attributed to its algorithmic synergy with the inherent physical properties of terahertz spectra. Firstly, THz spectra typically exhibit high multicollinearity due to the broadband nature of material absorption; unlike ordinary least squares, PLS effectively mitigates this by extracting a limited set of latent variables (LVs) that capture the maximal covariance between spectral features and material absorption. Secondly, PLS excels at dimensionality reduction, distilling relevant structure information while discarding redundant noise components. This mechanism inherently confers robustness against the high-frequency noise often present in DCT-transformed features, allowing PLS to maintain predictive accuracy where other algorithms struggle with overfitting to spectral artifacts.

**Figure 8.** The normalized performance metrics of the selected ML methods based on MSE, RMSE, MAE, R², t_{Train}, t_{Infer}, and t_{Total}.

RF ranked second, offering comparable accuracy but with a notably slower inference speed due to the computational overhead of aggregating multiple decision trees; nevertheless, its robustness to overfitting makes it a reliable non-linear alternative. In contrast, BP and SVM exhibited moderate accuracy but suffered from practical inefficiencies—BP's training time was prohibitively long for the achieved precision, while ELM and KNN performed poorly across all metrics. The failure of KNN, in particular, underscores the limitation of instance-based learning in high-dimensional spectral spaces corrupted by noise, where Euclidean distances become less meaningful, leading to large prediction variances.

The comparative evaluation of the six regression algorithms reveals a stark contrast in their suitability for THz-TDS thickness inversion. Synthesizing these results, PLS emerges as the unequivocal optimal model. Its superiority is attributed to its ability to extract latent variables from the DCT-compressed spectral features, effectively filtering the 20 dB Gaussian white noise while maximizing covariance with thickness. Conversely, KNN's failure highlights the curse of dimensionality in noisy spectral data, where local distance metrics break down.

On the other hand, it is important to evaluate the model's robustness under varying noise conditions. We compared the six candidate algorithms across a broad SNR range of 10~40 dB on MSE as summarized in Table 3, which serves as a reliable proxy due to its high covariance with established metrics such as RMSE, MAE, and R^2 in Figure 8. The results reveal distinct noise–response characteristics for each algorithm. RF demonstrates exceptional resilience, maintaining consistently low MSE values with minimal fluctuation across all noise levels, confirming its ensemble-based noise suppression capability. PLS regression exhibits a uniquely favorable profile that it achieves the lowest MSE among all methods at moderate-to-high SNRs. This superior performance highlights PLS's ability to exploit latent variable extraction for effective noise filtering once the signal becomes sufficiently discernible. In contrast, SVM displays a counter-intuitive trend, with MSE decreasing as noise increases, likely attributable to simplified decision boundaries that mitigate overfitting under low-SNR conditions. BP networks show steady performance improvements with increasing SNR, although they remain less efficient than PLS at higher SNRs. KNN proves highly sensitive to noise, with MSE peaking at 10 dB and only stabilizing at 40 dB, reflecting distorted distance metrics in noisy feature spaces. Most strikingly, ELM exhibits severe noise dependence that MSE escalates dramatically from 40 dB to 10 dB, underscoring the risks associated with minimal regularization and random projection mechanisms. Crucially, these results confirm that PLS and RF are the most robust candidates, with PLS offering superior accuracy in typical laboratory SNRs (≥ 20 dB) and RF providing unmatched stability in extremely noisy environments. These findings substantiate the generalizability of our conclusions beyond the initial 20 dB validation condition.

Table 3. The MSE comparison of the selected ML methods with 10~40 dB noise.

Noise	BP	SVM	RF	ELM	KNN	PLS
10 dB	0.099 ± 0.023	0.042 ± 0.002	0.009 ± 0.015	0.934 ± 0.024	0.805 ± 0.078	$0.0124 \pm 7 \times 10^{-4}$
20 dB	0.050 ± 0.008	0.032 ± 0.004	0.016 ± 0.024	0.209 ± 0.015	0.272 ± 0.016	$0.0021 \pm 2 \times 10^{-4}$
30 dB	0.017 ± 0.002	0.023 ± 0.002	0.021 ± 0.035	0.039 ± 0.004	0.077 ± 0.005	$1.4 \times 10^{-4} \pm 1 \times 10^{-5}$
40 dB	0.011 ± 0.003	0.142 ± 0.001	0.024 ± 0.042	$0.0050 \pm 4 \times 10^{-4}$	0.015 ± 0.001	$2.3 \times 10^{-5} \pm 2 \times 10^{-6}$

We also evaluate the model's robustness under different DCT numbers. Table 4 presents MSE of six machine learning methods under varying numbers of DCT coefficients 400~550 in a step of 50, the following trends can be observed. BP exhibits a gradual increase in MSE as the number of DCT coefficients grows, indicating a mild degradation

in predictive accuracy with higher-dimensional inputs. SVM shows a similar monotonic increase, suggesting that expanding the DCT feature set does not benefit SVM and may introduce noise or redundancy. RF behaves non-monotonically when MSE drops sharply from 400 to 450, then remains low for 500 and 550 coefficients. This implies that RF benefits from a moderate increase in DCT features but is relatively stable thereafter. ELM displays an unstable trajectory—starting with the lowest MSE at 400, deteriorating drastically at 500, and partially recovering at 550. The volatility reflects ELM’s sensitivity to input dimensionality and random initialization. KNN suffers a clear and consistent performance decline as DCT coefficients increase. This aligns with the known susceptibility of distance-based methods to the curse of dimensionality. PLS achieves the most favorable and consistently improving performance that MSE decreases steadily from 400 to 550. This indicates that PLS effectively exploits the expanded DCT representation, likely due to its intrinsic dimensionality reduction and alignment with latent variables. In summary, while most algorithms either degrade or fluctuate with more DCT coefficients, PLS demonstrates superior scalability and robustness, making it the most suitable choice for high-dimensional DCT-based regression in this context. RF also performs competitively after a threshold, whereas KNN and BP are adversely affected by feature expansion.

Table 4. The MSE comparison of the selected ML methods with different DCT numbers.

DCT No.	BP	SVM	RF	ELM	KNN	PLS
400	0.039 ± 0.007	0.023 ± 0.003	0.048 ± 0.098	0.021 ± 0.001	0.082 ± 0.005	$0.0324 \pm 1 \times 10^{-3}$
450	0.043 ± 0.008	0.026 ± 0.002	0.010 ± 0.009	0.070 ± 0.007	0.178 ± 0.012	$0.0327 \pm 7 \times 10^{-4}$
500	0.050 ± 0.008	0.032 ± 0.004	0.016 ± 0.024	0.209 ± 0.015	0.272 ± 0.016	$0.0021 \pm 2 \times 10^{-4}$
550	0.056 ± 0.012	0.036 ± 0.004	0.017 ± 0.029	0.045 ± 0.036	0.321 ± 0.026	$0.0014 \pm 1 \times 10^{-4}$

While nonlinear algorithms like BP and RF are theoretically capable of modeling complex mappings, their underperformance in this DCT-based context can be attributed to three key factors related to the nature of the transformed features. First, DCT coefficients are inherently highly collinear. The DCT rearranges the original spectral information into a descending order of energy compaction, meaning adjacent coefficients often represent similar frequency components. PLS is specifically designed to handle multicollinearity by projecting the predictors into a latent variable space that maximizes covariance with the response variable. Second, the curse of dimensionality impacts distance-based and tree-based methods differently than linear ones. Third, regarding the energy compaction property of the DCT, the lower-order coefficients primarily capture the global spectral trend, which often exhibits a quasi-linear relationship with the target variable. PLS excels at extracting these dominant latent structures. Nonlinear methods like BP, conversely, are prone to overfitting the high-frequency noise present in the higher-order coefficients if the regularization is not meticulously tuned—a challenge exacerbated by the limited sample size typical in engineering spectroscopy applications.

Therefore, PLS achieves the best performance not because the problem lacks nonlinearity, but because its intrinsic mechanism of latent variable extraction acts as a highly effective regularizer against the multicollinearity and high-dimensionality introduced by the DCT transformation.

Beyond the demonstrated accuracy, the computational efficiency of the proposed framework suggests potential scalability for handling larger datasets. While the current implementation exhibits low latency suitable for pilot-scale tests, further optimization will be pursued to ensure robust performance in data-intensive industrial environments. These

enhancements are expected to solidify the practical advantages of the selected algorithms for real-time deployment.

It is important to acknowledge that both the training and testing datasets were generated from the same simulation model. Consequently, the prediction task represents a relatively idealized scenario, which likely contributes to the uniformly high accuracy observed across all comparative models. In practice, however, a domain gap may exist between simulated and experimentally acquired THz signals due to factors such as system noise, environmental perturbations, and subtle variations in sample properties not captured by the model. Therefore, the reported performance metrics should be interpreted as an upper bound of the models' capabilities rather than definitive indicators of real-world efficacy. Bridging this simulation-to-reality gap through validation with experimental measurements remains a critical direction for future research.

4. Conclusions

This study systematically validated the efficacy of machine learning in overcoming the limitations of conventional THz-TDS thickness measurement methods. The paper successfully demonstrated a data-driven paradigm for THz film thickness measurement, effectively bypassing the convergence issues of model-based optimization and the interface-resolution limits of traditional ToF or model inversion methods. By leveraging FDTD-generated high-fidelity data, we demonstrated that data-driven models can accurately predict film thickness directly from noisy, compressed waveforms, eliminating the need for precise prior knowledge of optical constants required by model-based optimization. Among the tested algorithms, PLS stands out as the most effective solution, delivering unparalleled accuracy and real-time inference capabilities, making it ideally suited for industrial inline inspection. RF also proved to be a robust candidate for applications prioritizing maximum precision over processing speed. The remaining algorithms—BP, SVM, ELM, and KNN—failed to provide a competitive balance between precision and efficiency for this application. The outcome validates that machine learning, specifically PLS, can transform complex THz waveforms into precise thickness values suitable for real-time industrial monitoring.

It is important to note that this study relies exclusively on FDTD-simulated waveforms, while 20 dB Gaussian white noise was added to mimic realistic system SNR, real-world THz-TDS measurements introduce additional unmodeled uncertainties, including instrument timing jitter, alignment drifts, calibration errors, and batch-to-batch variability in material optical properties. These factors may degrade model performance relative to simulation results, though the inherent noise robustness of PLS and RF, combined with the DCT pre-processing pipeline, is expected to maintain acceptable accuracy in practical deployments. A second limitation of this work is the exclusive focus on a single film/substrate configuration with fixed optical properties. However, the proposed framework is inherently extensible to varied material systems for coatings with refractive indices ranging from 1.2 to 4.0 and extinction coefficients from 0 to 0.1, as well as diverse substrate materials. For multilayer film structures, additional reflected pulses from buried interfaces can be incorporated into the simulated waveforms, and the same DCT-based feature extraction and regression workflow applies directly. We will discuss the case for highly absorptive materials ($k > 0.1$) or ultra-low-index contrast systems ($\Delta n < 0.1$), which may increase prediction errors in future research.

Future work should focus on validating these findings with experimental THz measurements and exploring hybrid architectures that combine the physical constraints of wave propagation with the adaptive power of machine learning to further enhance generalization across diverse material systems. These findings lay the groundwork for embedding

such lightweight, high-precision algorithms into portable THz inspection devices, enabling rapid, non-destructive quality control for ultra-thin films in semiconductor manufacturing and flexible electronics, regardless of the operator's expertise in wave optics. In conclusion, PLS proved to be the most effective modeling approach for the simulated dataset presented herein. Nevertheless, we temper our conclusions regarding its immediate industrial application and real-time deployment, as these aspects require rigorous validation through experimental measurements. Such validation will be the focus of our subsequent research to bridge the gap between simulation and practice.

Author Contributions: Conceptualization, P.L. and L.Z.; methodology, X.L. and P.L.; software, X.L. and P.L.; validation, X.L., P.L. and Y.L.; formal analysis, X.L.; investigation, P.L. and Y.L.; resources, P.L.; data curation, X.L.; writing—original draft preparation, X.L.; writing—review and editing, P.L.; visualization, P.L.; supervision, P.L.; project administration, L.Z.; funding acquisition, L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (U2230128, 12204445, U2430203, 12427805); Presidential Foundation of CAEP (YZJJZC2022001).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are included in the article.

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

References

1. Park, J.; Cho, Y.J.; Chegal, W.; Lee, J.; Jang, Y.-S.; Jin, J. A review of thin-film thickness measurements using optical methods. *Int. J. Precis. Eng. Manuf.* **2024**, *25*, 1725–1737. [\[CrossRef\]](#)
2. Ming, Z.; Xiao, L.; Yang, L.; Ding, H.; Liu, J.; Yang, Z.; Liu, D.; Wang, K. Terahertz Nondestructive Measurement of Multilayer Films via Polarization Modeling Driven Deep Learning. *IEEE Sens. J.* **2025**, *25*, 20388–20395. [\[CrossRef\]](#)
3. Li, H.; Zhang, Z.; Nian, H.; Chen, Z.; Jiang, S.; Ding, F.; Sun, D.; Lin, H. Application of Terahertz Detection Technology in Non-Destructive Thickness Measurement. *Photonics* **2025**, *12*, 1191. [\[CrossRef\]](#)
4. Duvillaret, L.; Garet, F.; Coutaz, J.-L. A reliable method for extraction of material parameters in terahertz time-domain spectroscopy. *IEEE J. Sel. Top. Quantum Electron.* **1996**, *2*, 739–746. [\[CrossRef\]](#)
5. Wietzke, S.; Jansen, C.; Jung, T.; Reuter, M.; Baudrit, B.; Bastian, M.; Chatterjee, S.; Koch, M. Terahertz time-domain spectroscopy as a tool to monitor the glass transition in polymers. *Opt. Express* **2009**, *17*, 19006–19014. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Zhai, M.; Calvelli, S.; Shi, H.; Ricci, M.; Laureti, S.; Singh, P.; Fu, H.; Locquet, A.; Citrin, D.S. Comparative Study of Nondestructive Mapping of Conformal-Coating Thickness on Microelectronics by Terahertz Time-of-Flight Tomography. *IEEE Trans. Semicond. Manuf.* **2024**, *37*, 499–504. [\[CrossRef\]](#)
7. Zhong, S.C.; Du, L.H.; Jin, H.G.; Zhai, Z.H.; Li, J.; Zhu, L.G. Accurate thickness measurement method for thin layer in thick materials using terahertz technology. *Opt. Express* **2025**, *33*, 27626–27638. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Xu, Z.; Ye, D.; Chen, J.; Zhou, H. Novel terahertz nondestructive method for measuring the thickness of thin oxide scale using different hybrid machine learning models. *Coatings* **2020**, *10*, 805. [\[CrossRef\]](#)
9. Xu, Z.; Xu, J.; Wu, Y.; Yin, C.; Chen, S.; Liu, Q.; Ge, X.; Wan, L.; Ye, D. Classification and Prediction of Unknown Thermal Barrier Coating Thickness Based on Hybrid Machine Learning and Terahertz Nondestructive Characterization. *Coatings* **2025**, *15*, 725. [\[CrossRef\]](#)
10. Li, R.; Ye, D.; Xu, Z.; Zhou, H.; Yi, J.; Chen, Y.; Pan, J. Nondestructive evaluation of thermal barrier coatings thickness using terahertz time-domain spectroscopy combined with hybrid machine learning approaches. *Coatings* **2022**, *12*, 1875. [\[CrossRef\]](#)
11. Sun, F.; Cao, B.; Fan, M.; Ye, B.; Liu, L. THzMINet: A terahertz model-data-driven interpretable neural network for thickness measurement of thermal barrier coatings. *IEEE Trans. Ind. Inform.* **2023**, *20*, 4722–4734. [\[CrossRef\]](#)
12. Jiang, X.; Xu, Y.; Hu, H. Thickness characterization of steel plate coating materials with terahertz time-domain reflection spectroscopy based on BP neural network. *Sensors* **2024**, *24*, 4992. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Kaviya, P.P.; Arthitha, A.; Meenakshi, S.; Kumar, P.S. Advanced Non-Destructive Evaluation of Thermal Barrier Coatings Using THz-TDS and Machine Learning Techniques. In Proceedings of the 2025 IEEE International Conference on Contemporary Computing and Communications (InC4), Bangalore, India, 14–15 March 2025; pp. 1–6.

14. Wu, Y.; Cui, S.; Du, L.; Sarker, K.; Zhang, Z.; Li, J.; Zhang, M.; Yang, Z.; Li, L.; Wu, X. THz Automotive Paint Thickness Sensing Leveraging Physics-Informed Neural Network. *IEEE Trans. Terahertz Sci. Technol.* **2026**, *16*, 515–524. [[CrossRef](#)]
15. Pichler, M.; Hartig, F. Machine learning and deep learning—A review for ecologists. *Methods Ecol. Evol.* **2023**, *14*, 994–1016. [[CrossRef](#)]

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.