Data-Driven Optimization of TiO_2 Sol-Gel Synthesis: Insights from Statistical and Machine Learning Approaches

Gladys Egyir
Ohio University

Terfa Jude Igba
Georgetown University

Henry Makinde
University of North Carolina at Greensboro

Victor Stanley Francis
Georgetown University

Jeffrey Christian Ayerh
University of Ghana

Frederick Adrah
Ghana Communication Technology University

Dennis Opoku Boakye
University of North Carolina at Greensboro

ABSTRACT

Titanium dioxide (TiO_2) is used extensively in products from pigments and sunscreens to optical components. The sol–gel synthesis of TiO_2 is controlled by an intricate set of interactive parameters of which optimization is an important issue. A set of 290 experimental conditions was studied in detail to model and optimize yield of TiO_2 by means of statistical and machine learning methodologies. Out of the methodologies studied, polynomial regression and optimized random forest models showed best predictive capability achieving coefficient of determination (R^2) of 0.9522 and 0.9314, respectively, in comparison to linear regression. Feature importance analysis identified precursor concentration and hydrolysis ratio (water-to-precursor ratio) to play key role by having predominant influence, with secondary influence being aging time and pH. The paper highlights the value of data-based methodologies for synthesis design guidance, improved reproducibility, and expedited advances in materials chemistry.

General Terms

Machine Learning, Yield Prediction

Keywords

Machine Learning, Materials Synthesis

1. INTRODUCTION

Titanium dioxide (TiO_2) has been of great appeal as a multifunctional material of widespread use in future technologies. Its promise is made greater by nanometer-level engineering at which materials tend to exhibit new physicochemical properties due to the extraordinarily large ratio of volume-to-surface area and, on occasion, charge carrier quantum confinement [1][2].

Experiential performance of TiO_2 across its usages is inherently linked with its physicochemical characteristics, and these are governed by a complex balance of synthesis parameters. Detailed insight into how these dependent parameters affect yield and material quality is thereby critical for optimization of synthesis procedures and enabling production on a large scale [3]. As a result of significant developments in synthesis of TiO_2 using diverse routes, challenges persist due to the synergistic and sometimes non-linear character of these parameters. Conventional approaches, which tend to rely on iterative trial-and-error of the 'experimentation type', not only tend to be resource-intensive but also restricted in reliability [4].

Recent developments of data-intensive approaches provide a promising solution. The computational models, especially machine learning (ML) and statistical models, offer useful tools to reveal latent connections, discover patterns, and give prediction insights on materials synthesis and design [5][6][7].

Here, rigorous analysis of experimental dataset of TiO_2 synthesis conditions and yields are explored. The aim is twofold: (i) visualizing and describing interdependencies of important synthesis parameters and (ii) constructing prediction frameworks by using statistical and machine learning approaches. The framework takes a combination of scatter plot analysis, temporal trend visualization, ranking of feature importances and comparison of models on performance, and gives a system-level insight of the synthesis process.

2. RELATED WORK

A study presents the synergy of experimental design and machine learning methodologies for the optimization of catalyst synthesis. Specifically, the sol–gel conditions of a semi-hexagonal nanostructured calcium/titania–zirconia catalyst was modelled by multilayer perceptron and support vector machine models that exhibited high predictive capability. The calcination temperature was found to have the most significant influence, and optimization using genetic algorithm yielded catalysts with high surface area, well-defined nanoscale morphology, and good crystallinity. When optimized conditions were used, the catalyst reached 97.6% esterification conversion and showed steady performance for many cycles [8]

In a particular case study, machine learning methodologies were applied for predicting zinc oxide (ZnO) nanoparticle dimensions from synthesis conditions and band gap information using a sample set of 90 samples. Four individual ML models—i.e., CatBoost, Gradient Boosting, XGBoost, and a Stacking Ensemble—were developed, of which the Stacking Ensemble yielded the optimal level of precision ($R^2 = 0.9377$, MAE = 3.08 nm). A feature analysis indicated band gap as the most critical variable, and the model precisely predicted dimensions for unseen sets, matching well with experimental results achieved by scanning electron microscopy (SEM). A

graphical user interface was also generated that is easily interpretable, showcasing the potential of ML as a cost-effective and scalable means of predicting nanoparticle dimensions [9].

In a recent work, ML models were applied for the processing of thermogravimetric analysis (TGA) results acquired at different heating rates to predict and classify phase composition. A series of regression and classification algorithms, namely Gaussian Process Regression (GPR), k-Nearest Neighbor (KNN), Random Forest (RF), and XGBoost (XGB), were considered, and GPR achieved nearly perfect prediction accuracy ($R^2 = 0.999$) with small error margins. As for the classification performance, XGB achieved 99.9% accuracy, and RF and Decision Tree also showed excellent performance. The results demonstrate the potential of ML to optimize phase composition of TiO_2 nanomaterials efficiently and accurately and thereby shorten experimental times and computational costs [10].

3. PROBLEM STATEMENT

Given a dataset $\boldsymbol{X}=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\}$consisting of $\boldsymbol{N}$ samples, a machine learning model $\boldsymbol{f(x)}$ is employed to predict the yield of TiO_2 based on the input features.

4. METHODOLOGY

4.1 Synthesis of TiO_2

Titanium dioxide (TiO_2) was prepared through a controlled sol–gel procedure employing titanium alkoxide as the precursor. A set of Titanium alkoxide solutions were made in ethanol or isopropanol, hydrolyzed with deionized water at a specific H_2O-to-precursor mole ratio (usual range 4–20) with strong stirring. The reaction medium pH was set with dilute HCl or NH_4OH to reach acidic (pH ≈ 2) or basic conditions (pH ≈ 9–11), respectively, affecting particle size as well as gelation rate. The acquired sol was kept at 80 °C with aging time ranging from 2–48 h, during this time hydrolysis and condensation took place to produce TiO_2 nanostructures. The product-solid matter was centrifuged and successively washed with ethanol, as well as with de-ionized water, then heated to 100 °C to volatilize remaining solvents, with the final calcination carried out at 400–500 °C. The yield (%) was calculated from the ratio of the recovered TiO_2 mass to its initial experimental value.

4.2 Dataset overview

The dataset was aggregated from 290 separate sol–gel synthesis runs of TiO_2, each of which was described by numerous experimental parameters. Variables are explained and situated as follows:

Ti_alkoxide_molL: The molar concentration of the alkoxide precursor of the titanium, directly affecting nucleation and kinetics of growth. Higher precursor concentrations tend to favor greater particle formation but, if too high, will produce agglomeration and lower yield and reproducibility [11,12].

H_2O_to_precursor_ratio: The molar ratio of water to precursor applied to the hydrolysis and condensation reactions of the sol–gel process. A stoichiometric ratio ensures complete condensation and controlled network building, but deviations from stoichiometry (low or excess) create incomplete reaction or structural defects, negatively impacting yield [12,13,14].

Aging_time_hr: The time span of the aging period following synthesis, in which increase, and structural rearrangement takes place. Long aging times have the capability of boosting crystallinity and yield, but unduly lengthy times may lead to gel densification or unwanted transformations of phases, lessening efficiency [15,16].

pH: The acidity or alkalinity of the reaction medium, which controls substantially the kinetics of condensation and hydrolysis. The acidic pH slows down condensation and gives rise to particles of more uniform and of lower size, whereas alkaline pH increases gelation, and larger but less controllable structures tend to precipitate. The optimum pH conditions therefore decisively influence yield of TiO_2 [17,18].

Temperature: The reaction temperature, kept constant throughout all runs at 80 °C. While not a variable within this set of data, temperature control guarantees uniformity throughout experiment sets, as fluctuations would have a substantial impact on reaction rates and phase development [19].

Solvent: The ethanol or isopropanol solvent system, which affects precursor solubility and kinetics of reaction. The ethanol favors quicker hydrolysis on average, but by virtue of its less polar character, the isopropanol retards the reaction and provides divergent yield trends versus aging time [20,21,22].

Yield percent: The percentage of recovered TiO_2 from the synthesis expressed because of the experiment. The target variable for the machine and statistical models [23, 24].

4.3 Simulation Details

Preprocessing of the dataset was carried for the removal of missing and outlier values. The dataset was divided into test and training set. 70% of the dataset were used for training and 30% for test.

4.4 Exploratory Data Analysis

Scatter diagrams were generated to uncover the dependence of yield on three significant parameters: Ti_alkoxide_molL, H_2O_to_precursor_ratio, and pH. Aging trend analysis included plotting yield against time, stratified by solvent type, i.e., Isopropanol and ethanol. The aim of such plotting was to uncover solvent-based yield dependence on time.

4.5 Model building and Evaluation

Models were built to predict yield based on input attributes:

1. Linear Regression: A baseline of statistical models assuming linear relationships among features
2. Polynomial Regression: Nonlinear interactions and curvatures in the data are described.
3. Random Forest Regressor: An ensemble learning model that can learn complex, non-linear patterns and provide feature importance measures

Subsequently, hyperparameter optimization of the Random Forest model using GridSearch CV was done for optimal tree depth determination, optimum numbers of estimators, and minimum sample splits.

Models were evaluated using R^2 score, RMSE, and MAE. Feature importance was extracted from the tuned Random Forest Model.

5. DISCUSSION

In the following discussion, we combine observations from scatter plots, age trend examination, feature importance tanking, and comparison of machine learning models.

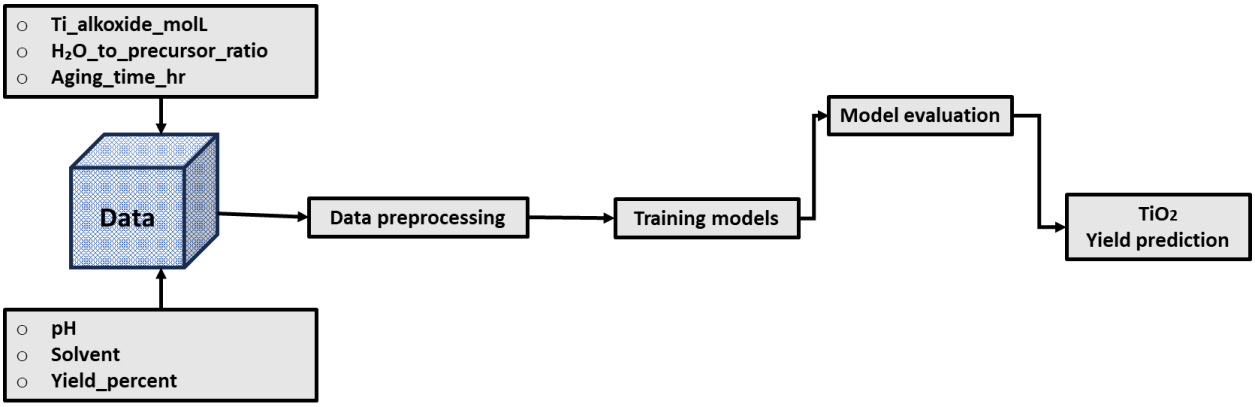

Fig 1: Flow chart for predicting yield of TiO_2

5.1 Multivariate Relationships and Scatter Plot Interpretation

Scatter plots of yield and synthesis parameters-showed non-linear and non-monotonic relationships. The following plots provide preliminary evidence that yield optimization is not feasible by means of single-variable tuning.

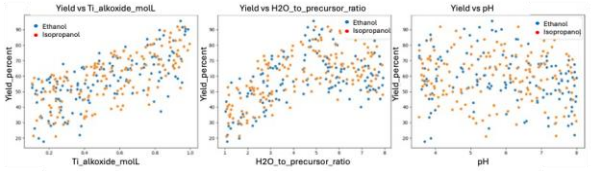

Fig 2: Yield vs (a) Ti_alkoxide_molL, (b H2O_to_precursor_ratio, and (c) pH

Yield vs Ti_alkoxide_molL: The yield reached its maximum at intermediate precursor concentrations meaning balance of having enough reactant availability without having enough to oversaturate and thereby lead to uncontrolled nucleation or particle agglomeration. This is consistent with principles of sol-gel chemistry wherein precursor concentration impacts rates of hydrolysis and condensation.

Yield vs H2O_to_precursor_ratio: From this scatter plot, it is shown that the water-to-precursor ratio has an intricate effect on yield. Relatively moderate ratios preferred larger yields, and high and low extremes were not favorable for yield. This may be due to incomplete hydrolysis at low ratios and dilution of intermediate species at high ratios.

Yield vs pH: From the scatter plot, it was observed that maximum yield was near neutral to weak conditions. Near strongly acidic conditions, there would be inhibition of hydrolysis and at extremes of base conditions, early condensation may get induced, and crystallinity may be low.

Deduction from Fig 2. reveal the need for multivariate optimization and propose that Titanium Alkoxide concentration (precursor) and hydrolysis ration are primary levers for improving and controlling yield.

5.2 Aging Trend Analysis

Preliminary analysis for aging trend, partly stratified by solvent type, if yield is influenced by aging time but with its impact mediated by the solvent environment. The systems based on ethanol showed more consistent yield increase with aging time, characteristic of a more controlled process of growth. The isopropanol systems showed more irregularity, which can arise from substantial differences in solvent polarity, solvent viscosity, or solvent-precursor interaction. Despite its chemical function in TiO_2 synthesis, aging time showed limited predictive relevance for the Random Forest model. The implication is that whereas aging may influence particle growth and crystallinity, its impact on yield is less consistent throughout the dataset and overridden by overriding factors.

Fig 3: Yield vs Aging time by solvent

5.3 Feature Importance from Tuned Random Forest Model

The noteworthy importance of the precursor concentration and water to precursor molar ratio helps to validate their critical role in yield determination. Such critical parameters significantly influence the hydrolysis and condensation reactions that give rise to TiO_2 particles. Conversely, the relatively low importance of pH, aging time, and solvent type suggests that the corresponding impacts bear a secondary or conditional nature and vary possibly with the concentrations of titanium alkoxide and the molar ratio of water to titanium alkoxide, which act as overwhelming determinants. The ranking provides a data-based approach to experimental variable prioritization in future synthesis efforts and challenges current assumptions on the effect of aging time on yield optimization.

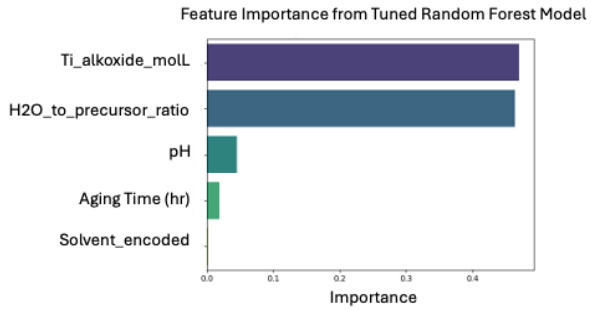

Fig 4: Feature importance from Tuned Random Forest Model

5.4 Residual Analysis and Model Diagnostics

The analysis of residuals of the Tuned Random Forest model showed a more symmetrical distribution centered at zero, which indicates low bias and a suitable model fit. In addition, the absence of skewness or heteroscedasticity adds strength to the model's validity for prediction use, showing that it generalizes well throughout the dataset.

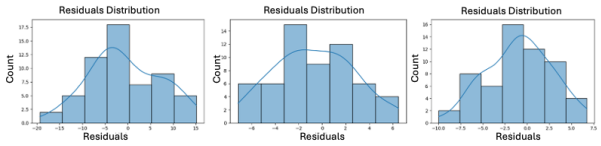

Fig 5: Residual analysis of (a) Linear regression, (b) Polynomial Regression, and (c) Tuned Random Forest

5.5 Model Comparison and Performance Evaluation

Linear regression served as a base method; nevertheless, it was not successful in well depicting the interactive and non-linear dynamics of the synthesis process. Polynomial regression achieved the largest value of R^2, indicating its improved capability of depicting curvature and feature interactions. The Tuned Random Forest model revealed robust performance, showing a relatively lower value of R^2 but providing improved interpretability and generalizability, which is strongly desirable for feature ranking and optimization.

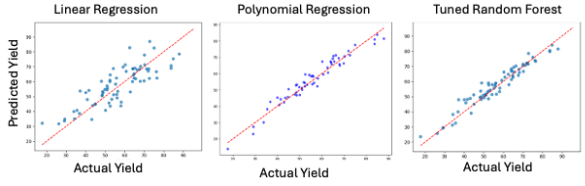

Fig 6: Comparison of Predicted and Actual Yields Across Models

Table 1.Model comparison and performance evaluation

Model	R^2 Score	RSME	MAE
Linear Regression	0.6932	7.8776	6.5165
Polynomial Regression	0.9426	3.4064	2.8232
Tuned Random Forest	0.9309	3.7394	2.9830

6. CONCLUSION

This research offers a data-intensive exploration of TiO_2 synthesis and shows that yield is managed through complex, non-linear interplay between numerous synthesis parameters. The most significant factors found to be influential were the mole ratio of water to precursor, titanium alkoxide concentration, pH, aging time, and solvent type. Of the models investigated, polynomial regression had the highest predictive power ($R^2 = 0.9426$), validating the inadequacy of simple linear models to describe the complexity inherent to the system. These results spotlight the power of combining experimental synthesis with state-of-the-art data analytics to improve material synthesis understanding and control. Extending beyond giving new insights into TiO_2 synthesis, this research lays the groundwork baseline knowledge of how artificial intelligence can be strategically brought to bear on the smart synthesis of nanomaterials, allowing for the enhancement of yield, optimization of reagents, and improvement of properties. Extensions of this framework can be made in future studies to procure and customize particle size, phase purity, and functional performance with the inclusion of additional variables such as dopants, surfactants, and thermal conditions. Eventually, this strategy is an early but crucial step toward intelligent, data-informed materials discovery as well as sustainable manufacturing.

7. REFERENCES

[1] Chen, X., & Selloni, A. (2014). Introduction: titanium dioxide (TiO2) nanomaterials. *Chemical reviews*, *114*(19), 9281-9282.

[2] Roduner, E. (2006). Size matters: why nanomaterials are different. *Chemical society reviews*, *35*(7), 583-592.

[3] Chen, X., & Mao, S. S. (2007). Titanium dioxide nanomaterials: synthesis, properties, modifications, and applications. *Chemical reviews*, *107*(7), 2891-2959.

[4] Liu, L., & Chen, X. (2014). Titanium dioxide nanomaterials: self-structural modifications. *Chemical reviews*, *114*(19), 9890-9918.

[5] Meuwly, M. (2021). Machine learning for chemical reactions. *Chemical Reviews*, *121*(16), 10218-10239.

[6] Adrah, F. A., Mottey, B. E., & Nyavor, H. (2024). The landscape of artificial intelligence applications in health information systems. *Int. J. Comput. Appl.*, *975*, 8887.

[7] Agboklu, M., Adrah, F. A., & Agbenyo, P. M. (2024). Hope Nyavor. From bits to atoms: machine learning and nanotechnology for cancer therapy. *Journal of Nanotechnology Research*, *6*, 16-26.

[8] Nayebzadeh, H., Rohani, A., Sistani, A., Hassanpour, A., & Gardy, J. (2022). Modelling and optimisation of the sol-gel conditions for synthesis of semi-hexagonal titania-based nano-catalyst for esterification reaction. *Catalysts*, *12*(2), 239.

[9] Alayou, S., Mengesha, M., & Tizazu, G. (2025). Application of machine learning models for predicting zinc oxide nanoparticle size. *Measurement*, 117785.

[10] Demirci, S., Şahin, D. Ö., & Demirci, S. (2025). Design of the amorphous/crystalline TiO2 nanocomposites via machine learning for photocatalytic applications. *Materials Science in Semiconductor Processing*, *192*, 109460.

[11] Karatchevtseva, I., Cassidy, D. J., Zhang, Z., Triani, G., Finnie, K. S., Cram, S. L., ... & Bartlett, J. R. (2008). Crystallization of TiO2 powders and thin films prepared from modified titanium alkoxide precursors. *Journal of the American Ceramic Society*, *91*(6), 2015-2023

[12] Hanaor, D. A., Chironi, I., Karatchevtseva, I., Triani, G., & Sorrell, C. C. (2012). Single and mixed phase TiO2 powders prepared by excess hydrolysis of titanium alkoxide. *Advances in Applied Ceramics*, *111*(3), 149-158.

[13] Guo, W., Lin, Z., Wang, X., & Song, G. (2003). Sonochemical synthesis of nanocrystalline TiO2 by hydrolysis of titanium alkoxides. *Microelectronic Engineering*, *66*(1-4), 95-101

[14] Karino, S., & Hojo, J. (2010). Synthesis and characterization of TiO2-coated SiO2 particles by hydrolysis of titanium alkoxide in alcohol solvents. *Journal of the Ceramic Society of Japan*, *118*(1379), 591-596.

[15] Jasbi, N. E., & Dorranian, D. (2016). Effect of aging on the properties of TiO2 nanoparticle. *Journal of Theoretical and Applied Physics*, *10*(3), 157-161.

[16] Hsiang, H. I., & Lin, S. C. (2004). Effects of aging on the phase transformation and sintering properties of TiO2 gels. *Materials Science and Engineering: A*, *380*(1-2), 67-72.

[17] Aida, I. S., & Sreekantan, S. (2011). Effect of pH on TiO2 nanoparticles via sol-gel method. *Advanced Materials Research*, *173*, 184-189.

[18] Yalcin, M. (2022). The effect of pH on the physical and structural properties of TiO2 nanoparticles. *Journal of Crystal Growth*, *585*, 126603.

[19] Chen, Y. F., Lee, C. Y., Yeng, M. Y., & Chiu, H. T. (2003). The effect of calcination temperature on the crystallinity of TiO2 nanopowders. *Journal of crystal growth*, *247*(3-4), 363-370

[20] Hu, L., Yoko, T., Kozuka, H., & Sakka, S. (1992). Effects of solvent on properties of sol—gel-derived TiO2 coating films. *Thin solid films*, *219*(1-2), 18-23.

[21] Rao, Y., Antalek, B., Minter, J., Mourey, T., Blanton, T., Slater, G., ... & Fornalik, J. (2009). Organic solvent-dispersed TiO2 nanoparticle characterization. *Langmuir*, *25*(21), 12713-12720.

[22] Kojima, T., & Sugimoto, T. (2008). Formation mechanism of amorphous TiO2 spheres in organic solvents 3. Effects of water, temperature, and solvent composition. *The Journal of Physical Chemistry C*, *112*(47), 18445-18454.

[23] Mohammadi, S., Harvey, A., & Boodhoo, K. V. (2014). Synthesis of TiO2 nanoparticles in a spinning disc reactor. *Chemical Engineering Journal*, *258*, 171-184.

[24] Adrah, F. A., Denu, M. K., & Buadu, M. A. E. (2023). Nanotechnology applications in healthcare with emphasis on sustainable COVID-19 management. *J Nanotechnol Res*, *5*, 6-13.