Scientific African 9 (2020) e00516

Contents lists available at ScienceDirect

Scientific African

journal homepage: www.elsevier.com/locate/sciaf

Ensemble learning prediction of transmittance at different wavenumbers in natural hydroxyapatite☆

Emmanuel Okafor[a,*], David O. Obada[b,d], David Dodoo-Arhin[c]

[a] Department of Computer Engineering, Ahmadu Bello University, Zaria, Nigeria
[b] Department of Mechanical Engineering, Ahmadu Bello University, Zaria, Nigeria
[c] Department of Materials Science and Engineering, University of Ghana, Accra, Ghana
[d] Africa Center of Excellence on New Pedagogies in Engineering Education, Ahmadu Bello University, Zaria, Nigeria

ARTICLE INFO

Article history:
Received 4 March 2020
Revised 31 July 2020
Accepted 2 August 2020

Keywords:
Ensemble learning
Transmittance prediction
Hydroxyapatite
Nucleating sites

ABSTRACT

Material engineering-based research has often relied so much on tedious human experiments for generating specific engineering properties with a major draw-back of high time demand that can span between an hour and days. Hence to deviate from the usual paradigm, we provide an alternative approach which employs artificial intelligence (AI) based ensemble learning methods for predicting the degree of transmittance for a range of wavenumbers of infrared radiation through hydroxyapatite (HAp) samples. The effective samples (transmittance and wavenumber) were passed as input to the predictive systems. For this, we trained two ensemble learning methods: E**x**treme **G**radient **B**oosting (XGBoost) and Random Forest on variants of HAp (density and time variations), while considering a fixed amount of 10,000 base estimators. The results show that Random Forest marginally outperforms the XGBoost in the testing phase but requires a much longer computing time. However, XGBoost is much faster than the Random Forest. Furthermore, the examined ensemble learning models yielded a coefficient of determination ($R^2 > 0.997$): which are in close agreement with experimental data, depicting an excellent generalization capacity. Additionally, the examined ensemble learning models showed a significant $\geq$ 99.83% decrease in computational complexity relative to the time spent when generating the experimental data. Overall, the use of ensemble learning models is very important for validating material engineering properties.

1. Introduction

Hydroxyapatite (HAp) is a calcium phosphate, with chemical formula $Ca_{10}(PO_4)_6(OH)_2$ which typically exists as the main material of bones and teeth with high biaffinity. The functional groups on the surface of biomaterials like hydroxyapatite play a very important role in nucleating calcium phosphate deposition on surgical implants. In the context of bone implants

☆ Editor: Bolanle Adefowoke Ojokoh

* Corresponding author.
E-mail addresses: e.okaforc@gmail.com, eokafor@abu.edu.ng (E. Okafor).

https://doi.org/10.1016/j.sciaf.2020.e00516

for load bearing application, ceramic implants are preferred to their metallic counterparts as some metals and its alloys, for instance, Titanium (*Ti*) alloys, do not bond to bone in post-implantation stage [1]. Because of the limitation on metal-based implants, several methods have been employed to embed hydroxyapatite on the surfaces of metal implants to enhance the integration with bony tissues. The central goal of this study is to employ artificial intelligence-based supervised learning algorithms for understanding the underlying non-linear patterns of transmittance at different wavenumbers of infrared radiation through the bulk of the hydroxyapatite specimens.

Based on the motivation from the stated goal, it is pertinent to understand the concept of biomineralization, a naturally driven process, which involves the initialization and formation of bones and teeth and has been used to produce hydroxyapatite coatings. In this method, bioactive apatite layers are formed on a support/substrate by immersing in synthesized body fluid, for instance, simulated body fluid (SBF) which has a very close similarity to the human body fluid. A key step in the aforementioned coating process is the nucleation of calcium phosphate on the surface of the implant which is highly influenced by the functional groups of the coating material.

Early studies have confirmed the role of functional groups in inducing early nucleation of calcium phosphates on the surface of metal alloy substrates. The research work by Tanahashi and Matsuda [2] formed a monolayer on gold substrates using alkanethiols to investigate the surface functional groups on the efficiency to nucleate calcium phosphates after immersion in SBF. Their research reported that the most efficient nucleating group was -PO_4H_2. The research paper by Mao et al. [3] introduced carboxyl and hydroxyl groups on the surface of titanium-based implants and the introduced functional groups revealed the nucleation of apatite. These outcomes confirm that functional groups vis-a-vis their transmittances at different wavelengths have a central role to play in nucleation. Hydroxyapatite (HAp) has shown very efficient nucleation sites and serves as an enhancer for the formation of apatite. This is because HAp has a very close structural similarity with the human bone in addition to increased concentration of calcium and phosphate when HAp dissolves in the surrounding medium [4,5].

For more than a decade now, the use of machine learning algorithms [6] have aided in the modeling of linear or non-linear interpretation of a given data. Intelligent computational techniques have shown a strong potential to learn the non-linear relationships by using historical information from experimental datasets. These obtained models from learning techniques can significantly lower the testing times involved in the fabrication of new materials in addition to their characterization. Some studies have shown how artificial intelligence has been used to predict material properties such as wear [7], nonlinear dynamic mechanical properties [8], tensile strength [9], and erosive wear [10]. We are of the opinion that voluminous laboratory analyses can be streamlined by using computational methods.

Contribution: Previous research has often discussed material composition, nucleation properties, and the use of neural networks for predicting specific properties of engineering materials. In this study, we report the use of ensemble learning methods for predicting the transmittance levels for a given wavenumber of infrared radiation through the bulk of HAp specimens. The motivation for the choice of ensemble learning technique stems from a general opinion that most ensemble learning systems have proven to be effective and extremely versatile in a broad spectrum of problem domains [11]. We found that ensemble learning techniques present a coefficient of determination ($R^2 > 0.997$), indicating an excellent generalization capacity because the model results were in close agreement with the laboratory experimental data. Moreover, the used ensemble learning methods present computational benefits as against laboratory experimental rigor. Additionally, the study demonstrated that in the testing phase, the Random Forest marginally outperforms the XGBoost on all variants of the used data but exhibits a drawback of higher computational complexity with respect to time.

2. Data curation

Non-separated animal bones which are regarded as wastes were obtained from an abattoir in Zaria, Nigeria, the sourced bones (raw materials) were used as precursors for hydroxyapatite production. The sourced bones were thoroughly cleaned and a detailed procedure can be found in studies by our group [12,13]. Tap water was used for cleaning the bones throughout the process. The cleaned bones were subjected to sintering at 900 °C for 2 h, 4 h and 6 h at a heating rate of 5 °C/min. After sintering, the obtained HAp powders were milled in a metallic mortar and passed through a sieve of 300 μm mesh prior to powder analyses.

2.1. Fourier transform infrared spectroscopy

To determine the functional group present in the raw materials and equally, the extent of probable de-hydroxylation during the heat treatment, Fourier Transform Infrared (FT-IR) spectroscopy data were collected on an FT-IR spectrometer equipped with Universal Attenuated Total Reflection (UATR) sampling accessory. The change in the functional group of the raw materials was verified in the wavenumber range of 648–4000 cm^{-1}. The HAp specimens were grinded and mixed with dried potassium bromide (KBr) using a metallic mortar and loaded into a sample holder mounted in the instrument.

2.2. Data distribution

The data variants contain 900 examples (wavenumber(X) and transmittance (Y)) within the data lexicon. In Table 1, we summarize three different dataset description. For example in data HA-2 h, the input feature (X_a is a one dimensional

Table 1
Dataset variants containing a pair of transmittance and different wavenumber values.

S/N	HA-2 h		HA-4 h		HA-6 h	
	X_a	Y_a	X_b	Y_b	X_c	Y_c
1	3999.434	98.688	3999.434	98.648	3999.434	98.856
2	3995.706	98.667	3995.706	98.655	3995.706	98.803
3	3991.979	98.686	3991.979	98.637	3991.979	98.810
...	...	...	...	...	...	...
898	656.011	87.667	656.011	88.147	656.011	90.864
899	652.284	85.753	652.284	86.525	652.284	88.822
900	648.557	82.780	648.557	83.859	648.557	86.444

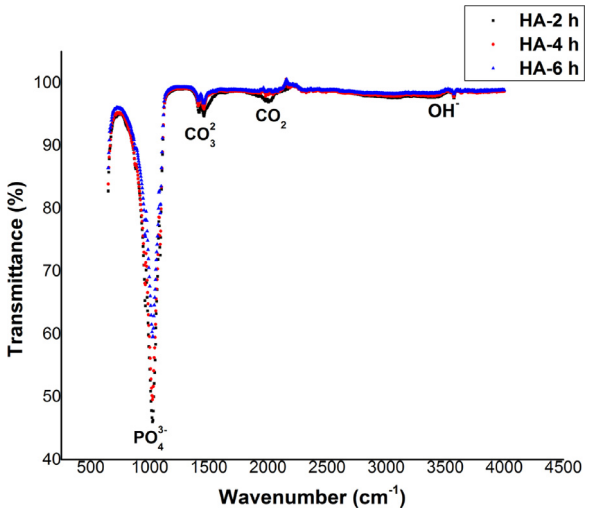


Fig. 1. Illustration of the HAp data lexicons showing the wavenumber vs the transmittance and the functional groups within the bulk of the hydroxyapatite specimens. Each of the HAps was subjected to a fixed sintering temperature of 900°C over a varying time duration of 2 h, 4 h, and 6 h .

feature vector) and the ground-truth output value (Y_a) from normal engineering process represent the wavenumber and transmittance levels respectively in the bulk-specimen of the natural hydroxyapatite. The table also describes the feature information for the other datasets (HA-4 h and HA-6 h). The graphical illustration of the variants of data distributions is shown in Fig. 1.

We partitioned each of the dataset variants into the ratio 80%: 20% for the training set and testing set respectively. Note that, the data partitioning were randomly shuffled five possible times, while ensuring that the different testing set examples do not overlap within the training set examples.

3. Supervised learning methods

In this section, two ensemble methods were investigated;

3.1. Random forest

Random Forest [14] is an ensemble learning technique derived from bagging aggregation. This is often done by building the integration of several de-correlated trees. After which an effective average of the base estimators is taken. We adopted similar Random Forest modeling procedures as described by Biau [15].

Given that the training example $D_m = (X_1, Y_1), (X_2, Y_2), \ldots (X_m, Y_m)$ of independent and identically distributed (i.i.d). The $[0, 1]^k \times$ R-valued random variables with similar distribution as an independent set (X_a, Y_a). For a fixed $x \in [0, 1]^k$, our objective is to compute a regression function $r(x) = E[Y|X = x]$ from a given data lexicon D_m. A Random Forest can be described as a predictive estimator that consist of a collection set of randomized base regression trees $\{r_m(x, \theta_k, D_m), k \geq 2\}$, where the variable $\theta = \{\theta_1, \theta_2, \ldots\}$ are randomly distributed variable. The random output descision trees were integrated for creation of aggregated regression trees;

$$\bar{r_m}(X, D_m) = E_\theta[r_m(X, \theta_k, D_m)] \tag{1}$$

E_θ denote expectation with respect to the random parameter conditionally on X and the data set D_m. We computed $\bar{r_m}(X)$ instead of $\bar{r_m}(X, D_m)$ based on a Monte Carlo principle, by generating N_{est} random trees and computing an average of the

Table 2
XGBoost hyperparameters.

Hyperparameters	L_r	N_{est}	$D_{\max}$	H_{jm}	γ	S_{sub}	S_{col}	α_R	λ_R
Initialization Values	0.07	10,000	3.0	1.5	0	0.6	0.4	0.75	0.45

individual outcomes from the predictive estimators. For this, we employed ($N_{est} = 10,000$) base estimators. The two main steps required during random forest estimation;

- At each node, present a coordinate $X = (X^{(1)}, X^{(2)} \ldots X^{(k)})$ is selected, whereby the jth feature for a given probability $Pr_{mj} \in (0, 1)$.
- In a scenario whereby a coordinate is selected, a split would be created at the mid-point of the selected side.

A compact expression of $\bar{r_m}(X)$ estimation [15] can be defined as;

$$\bar{r_m}(X) = E_\theta[r_m(X, \theta)] = E_\theta\left[\frac{\sum_{i=1}^{m} Y_i 1_{[X_i \in A_m(X,\theta)]}}{\sum_{i=1}^{m} 1_{[X_i \in A_m(X,\theta)]}} 1_{E_m(X,\theta)}\right] \tag{2}$$

where $A_m(X, \theta)$ accounts for the rectangular cell for the random partitioning that contains X.

3.2. Extreme gradient boosting (XGBoost)

The XGBoost [16] is a gradient boosting method which employs cache access pattern, sharding, and data-compression for building a more scalable ensemble technique (tree boosting system). This technique employs sparsity awareness and weighted quantile sketch algorithms. The weighted quantile sketch algorithm approximates the tree learning and the sparsity awareness algorithm handles sparse data features. This method processes the input features and then constructs a model that can perform any of the following tasks; classification, regression, and ranking. The XGBoost was mainly used for regressive prediction purposes.

A compact mathematical formulation of the objective function[1] in XGBoost can be defined as;

$$L = \sum_{j=1}^{T_m}[G_{jm} w_{jm} + \frac{1}{2}(H_{jm} + \lambda_R) w_{jm}^2 + \alpha_R |w_{jm}|] + \gamma T_m. \tag{3}$$

where L is the objective function, α_R and λ_R denote L_1 and L_2 regularization term respectively, γ denote the penalization term also known as lagrangian multiplier, it is mainly used for controlling complexity during splitting. The H_{jm} and G_{jm} denote sum of hessian and sum of gradient respectively for the optimal weight for each region j, (H_{jm} was used to determine the minimum sum of child weights) and w_{jm} denote the optimal weights. T_m represents leave tree over a maximum depth of the tree $D_{\max}$. Note that the variables; H_{jm} and G_{jm} are dependent on the input variable X, from the previously described data D_m. Further explanation of the colsample by tree level (S_{col}), subsample (S_{sub}), sparse-aware, and shrinkage information were discussed in paper by Chen and Guestrin [16]. In the experiment, we considered manual tuning of the XGBoost hyperparameters. The summary of the hyperparameters of the XGBoost were reported in Table 2.

Experimental Setup: we employed python programming language for implementing each of the predictive models. The computing resource used during training of the predictive model: An ASUS Sonic Master laptop which consists of the following hardware specifications; 8 GB of RAM, 2.0 GB of NVIDIA GPU, and CPU (CORE i7).

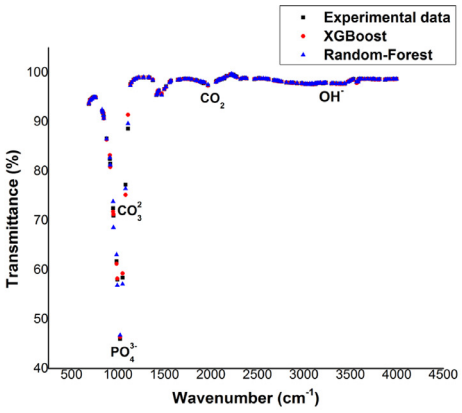
4. Results

This section entails the discussion of the performance metrics of the ensemble learning methods; which is reported in the context of the error metrics and coefficient of determination (R^2). A model is said to have good predictive properties when the mean absolute error (MAE) and root mean square error (RMSE) is very low with a corresponding high $R^2 > 0.5$.

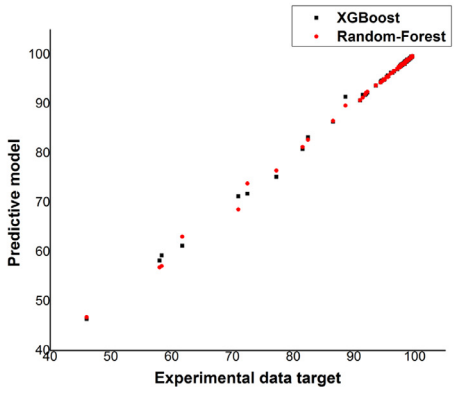
4.1. Results analysis

We report the visualization of the ensemble learning methods as shown in Fig. 2. From this figure, the learning curves and the correlation plot from both methods, suggest that these techniques exhibit an almost similar level of generalization potential; the conclusion stems from the observation whereby the predicted models and experimental data are highly correlated. Note that the two ensemble learning methods were trained based on five-fold cross-validation, then we computed an effective mean of the error metrics and R^2 with their corresponding standard deviations. The summary of the performance measures of the two methods in the training phase was reported in Table 3. From this table, we observed that the XGBoost yielded a lower error metric compared with Random-Forest on all the used data variants. In the testing phase, the Random

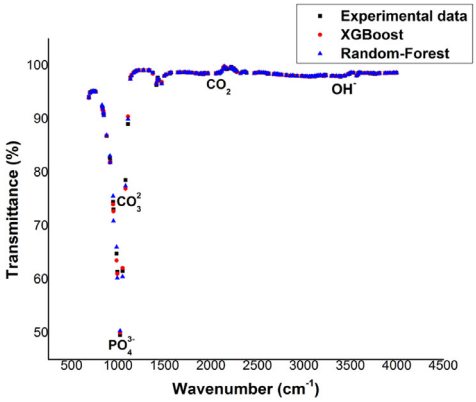
[1] https://towardsdatascience.com/boosting-algorithm-xgboost-4d9ec0207d



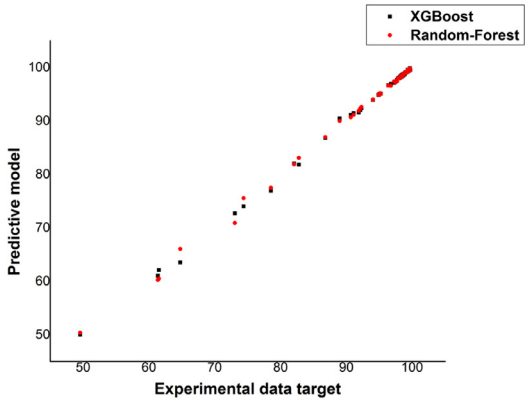
(a) HAp after 2h sintering dwell time (HA-2h)



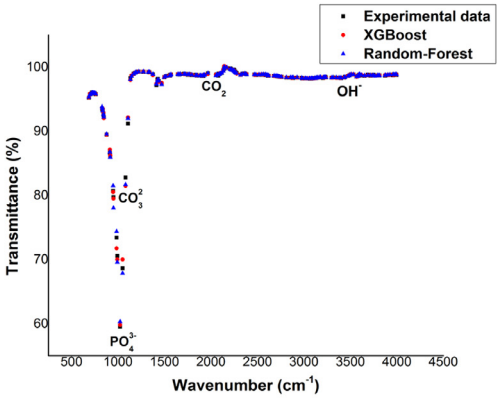
(b) HA-2h



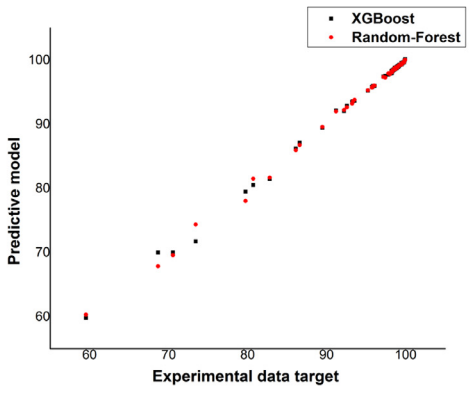
(c) HAp after 4h sintering dwell time (HA-4h)



(d) HA-4h



(e) HAp after 6h sintering dwell time (HA-6h)



(f) HA-6h

Fig. 2. Testing phase performance visualization of the ensemble learning models on different variants of the HAp data (4th experimental run): all the left hand plots represent the learning curve for each of the examined ensemble techniques and all the right-hand plots illustrate the correlation between the target value (experimental data generated from normal engineering process) and the predictive values from the ensemble learning techniques. Note that each of the testing set contains 180 examples (target) and predictive values were represented in the dot form distributions.

Table 3
Training phase performance metric evaluation of the ensemble learning methods on variants of the used dataset after five-folds cross-validation (metric mean plus standard deviation).

Data variants	Learning methods	MAE	RMSE	R^2
HA-2 h	Random Forest	$0.0489 \pm 7.85 \times 10^{-4}$	$0.1534 \pm 1.15 \times 10^{-2}$	$0.9996 \pm 8.21 \times 10^{-5}$
	XGBoost	$0.0397 \pm 1.45 \times 10^{-3}$	$0.1076 \pm 1.92 \times 10^{-2}$	$0.9998 \pm 8.75 \times 10^{-5}$
HA-4 h	Random Forest	$0.0472 \pm 8.93 \times 10^{-4}$	$0.1407 \pm 9.26 \times 10^{-3}$	$0.9996 \pm 7.17 \times 10^{-5}$
	XGboost	$0.0412 \pm 9.46 \times 10^{-4}$	$0.0997 \pm 1.37 \times 10^{-2}$	$0.9998 \pm 6.76 \times 10^{-5}$
HA-6 h	Random Forest	$0.0391 \pm 8.54 \times 10^{-4}$	$0.1205 \pm 8.27 \times 10^{-3}$	$0.9996 \pm 8.91 \times 10^{-5}$
	XGboost	$0.0387 \pm 1.15 \times 10^{-3}$	$0.0938 \pm 1.77 \times 10^{-2}$	$0.9997 \pm 1.31 \times 10^{-4}$

Table 4
Testing phase performance metric evaluation of the ensemble learning methods on variants of the used dataset after five-folds cross-validation (metric mean plus standard deviation).

Data variants	Learning methods	MAE	RMSE	R^2
HA-2 h	Random Forest	$0.1166 \pm 1.64 \times 10^{-2}$	$0.3273 \pm 6.07 \times 10^{-2}$	$0.9983 \pm 2.91 \times 10^{-4}$
	XGBoost	$0.1374 \pm 3.23 \times 10^{-2}$	$0.4035 \pm 1.10 \times 10^{-1}$	$0.9974 \pm 7.89 \times 10^{-4}$
HA-4h	Random Forest	$0.1126 \pm 1.54 \times 10^{-2}$	$0.2989 \pm 6.04 \times 10^{-2}$	$0.9983 \pm 3.33 \times 10^{-4}$
	XGboost	$0.1311 \pm 2.95 \times 10^{-2}$	$0.3623 \pm 1.11 \times 10^{-1}$	$0.9976 \pm 9.31 \times 10^{-4}$
HA-6h	Random Forest	$0.0961 \pm 1.39 \times 10^{-2}$	$0.2515 \pm 4.64 \times 10^{-2}$	$0.9981 \pm 2.94 \times 10^{-4}$
	XGboost	$0.1111 \pm 1.93 \times 10^{-2}$	$0.3031 \pm 7.46 \times 10^{-2}$	$0.9972 \pm 7.09 \times 10^{-4}$

Table 5
Mean time analysis of the ensemble learning models and the normal engineering process (NEP).

Data variants	NEP	Random forest	XGBoost
HA-2 h	2 h (7200 s)	11.94 s (99.83% decrease)	3.39 s (99.95% decrease)
HA-4 h	4 h (14,400 s)	11.86 s (99.92% decrease)	3.28 s (99.98% decrease)
HA-6 h	6 h (21,600 s)	11.80 s (99.94% decrease)	3.12 s (99.99% decrease)

Forest obtained a higher mean coefficient of determination ($R^2 \geq 0.9981$) and minimal error metrics (MAE and RMSE) when compared with XGBoost which yielded $R^2 = 0.9976$ on a more dense HAp (HA-4 h).

The summary of the results obtained in the testing phase was reported in Table 4. Note that for each run of the experiments, we report that the summation of the training and testing time for all the examined ensemble learning methods were approximately $t < 12$ s, a summary of the mean computational complexity analysis for each of the methods is summarized in Table 5. From the described table, we report that the ensemble learning models exhibit computing time decay on a more denser variant of the used data. The XGBoost showed $\times$ 2.52 faster computing time compared with Random Forest. Moreover, the examined ensemble learning techniques showed a $\geq$ 99.83% decrease in computational complexity when compared with the time spent for generating the HAp from the normal engineering process (NEP).

We report that in the training phase, the ensemble learning techniques yielded lower MAE and better performance stability relative to the testing phase (the ensemble learning obtained higher MAE). The reason for the lower error metrics (MAE/RMSE) in the training phase could have arisen because the ensemble learning algorithms were trained on 80% of the entire data, in each dataset. Hence the resulting trained models contained more informative knowledge about the used data lexicons than during the testing phase evaluation. Overall the error levels of the ensemble learning model in both evaluation phases (training and testing) are considerably low ($MAE \leq 0.14$). Although the comparison of the coefficient of determination ($R^2 \geq 0.997$) in both evaluation phases suggest that, the ensemble learning models depict excellent generalization and are very correlated relative to the NEP data.

5. Conclusion

This research has compared two forms of ensemble learning techniques (XGBoost and Random Forest) which were used for predicting the transmittance levels for a given wavenumber of infrared radiation through a variant of bulk HAp specimens. The generated models were able to accurately reveal the degree of correlation between the target and predicted transmittance levels within the lattice of the HAp samples. We found that ensemble learning techniques present a coefficient of determination ($R^2 > 0.997$), indicating an excellent generalization capacity because the model results were in close agreement with the laboratory experimental data. Moreover, the used ensemble learning methods present computational benefits as against laboratory experimental rigor. Additionally, the study demonstrated that in the testing phase, the Random Forest marginally outperforms the XGBoost on all variants of the used data but exhibits a drawback of higher computational complexity with respect to time. Furthermore, the study demonstrated the relevance of using artificial intelligence models for validating engineering materials and provides an advantage in generating well-correlated models with very low

mean computing time ($t < 12$ s) as against human engineering rigor that could span for hours. Future work can compare recurrent neural networks with the examined techniques, to assess the possibility to improve the current performance.

Acknowledgements

The authors acknowledge the Department of Computer Engineering and Department of Mechanical Engineering, Ahmadu Bello University, Zaria, Nigeria for providing facilities to carry out this study. DDA acknowledges support from the University of Ghana BANGA-Africa program.

Funding

The work received no funding from external sources.

Declaration of Competing Interest

The authors declares no potential conflict of interests.

References

[1] Q. Liu, J. Ding, F.K. Mante, S.L. Wunder, G.R. Baran, The role of surface functional groups in calcium phosphate nucleation on titanium foil: a self--assembled monolayer technique, Biomaterials 23 (15) (2002) 3103–3111.
[2] M. Tanahashi, T. Matsuda, Surface functional group dependence on apatite formation on self-assembled monolayers in a simulated body fluid, J. Biomed. Mater. Res. 34 (3) (1997) 305–315.
[3] C. Mao, H. Li, F. Cui, Q. Feng, H. Wang, C. Ma, Oriented growth of hydroxyapatite on (0001) textured titanium with functionalized self-assembled silane monolayer as template, J. Mater. Chem. 8 (12) (1998) 2795–2801.
[4] T.P. Coogan, D.M. Latta, E.T. Snow, M. Costa, A. Lawrence, Toxicity and carcinogenicity of nickel compounds, CRC Crit. Rev. Toxicol. 19 (4) (1989) 341–384.
[5] K. Alvarez, K. Sato, S. Hyun, H. Nakajima, Fabrication and properties of lotus-type porous nickel-free stainless steel for biomedical applications, Mater. Sci. Eng. 28 (1) (2008) 44–50.
[6] K. Hasegawa, K. Funatsu, Non-linear modeling and chemical interpretation with aid of support vector machine and regression, Curr. Comput. Aided Drug Des. 6 (1) (2010) 24–36.
[7] Z. Jiang, Z. Zhang, K. Friedrich, Prediction on wear properties of polymer composites with artificial neural networks, Compos. Sci. Technol. 67 (2) (2007) 168–176.
[8] M. Al-Haik, M. Hussaini, H. Garmestani, Prediction of nonlinear viscoelastic behavior of polymeric composites using an artificial neural network, Int. J. Plast. 22 (7) (2006) 1367–1392.
[9] M. Trebar, Z. Susteric, U. Lotric, Predicting mechanical properties of elastomers with neural networks, Polymer 48 (18) (2007) 5340–5347.
[10] Z. Zhang, N.-M. Barkoula, J. Karger-Kocsis, K. Friedrich, Artificial neural network predictions on erosive wear of polymers, Wear 255 (1–6) (2003) 708–713.
[11] R. Polikar, Ensemble learning, in: Ensemble Machine Learning, Springer, 2012, pp. 1–34.
[12] D. Obada, E. Dauda, J. Abifarin, D. Dodoo-Arhin, N. Bansod, Mechanical properties of natural hydroxyapatite using low cold compaction pressure: effect of sintering temperature, Mater. Chem. Phys. 239 (2020) 122099.
[13] J. Abifarin, D. Obada, E. Dauda, D. Dodoo-Arhin, Experimental data on the characterization of hydroxyapatite synthesized from biowastes, Data Brief 26 (2019) 104485.
[14] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[15] G. Biau, Analysis of a random forests model, J. Mach. Learn. Res. 13 (1) (2012) 1063–1095.
[16] T. Chen, C. Guestrin, XGboost: a scalable tree boosting system, in: Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 785–794.