

RESEARCH ARTICLE

WILEY

Prediction of the reflection intensity of natural hydroxyapatite using generalized linear model and ensemble learning methods

Emmanuel Okafor¹  | David O. Obada^{2,3} | Yusuf Ibrahim¹ | David Dodoo-Arhin⁴ 

¹Department of Computer Engineering,
Ahmadu Bello University, Zaria, Nigeria

²Department of Mechanical Engineering,
Ahmadu Bello University, Zaria, Nigeria

³Africa Center of Excellence on New
Pedagogies in Engineering Education,
Ahmadu Bello University, Zaria, Nigeria

⁴Department of Materials Science and
Engineering, University of Ghana, Accra,
Ghana

Correspondence

David Dodoo-Arhin, Department of
Materials Science and Engineering,
University of Ghana, Accra 77, Ghana.
Email: ddarhin@yahoo.com

Abstract

Laboratory data acquisition and analysis of X-ray diffraction (XRD) data involves a lot of tedious human engineering and is time-consuming. To put in context, a summation of the material synthesis procedure leading to the analysis of the structure of the material can span several days. To curb this challenge and to enhance innovations in engineering pedagogy, this article investigates an alternative method that uses supervised learning algorithms based on ensemble techniques and a generalized linear model (GLM) for predicting reflection intensity (XRD patterns) of raw and natural hydroxyapatite under varying sintering temperature conditions given Bragg angles as input to the machine learning algorithms. For the experiment, we trained GLM and ensemble learning models (CatBoost, LightGBM, and two variants of XGBoost based on manual and genetic algorithm for tuning of the hyperparameters). The results show that most instances of the XGBoost yielded a robust performance that surpasses all other approaches, when predicting X-ray reflection intensities ascribed to the biomaterials subjected to varying sintering temperature conditions. In addition, the results show that all the ensemble techniques significantly outperform the GLM, this indicates that the former exhibits better generalization capacity. The ensemble learning techniques and the GLM presents a reduced computational complexity.

KEYWORDS

biomaterials, ensemble learning, generalized linear model, predictive modeling

1 | INTRODUCTION

Biomaterials play an important role in the field of materials technology and biomedical engineering. This is because hydroxyapatite (HAp), a prominent calcium-phosphate finds practical application in orthopedics and dentistry¹ making the crystallinity and of course, other properties of HAp very important features for its medical application. Most materials have characteristic reflections at certain Bragg angles, which often reveal the identity of a material at a given point in

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *Engineering Reports* published by John Wiley & Sons Ltd.

time. Hence, it is pertinent to understand the underlying mapping between Bragg angles (of raw and HAp powders under varying sintering temperature conditions) and the target output (reflection intensity in counts). A broad focus of this study is aimed at training artificial intelligent (AI) models that can draw inference knowledge from the reflection intensities (RI) given Bragg angles of raw and synthesized HAp powders as input to the supervised learning techniques.

Early research in the field of AI has employed pattern recognition and artificial neural networks (ANN) for constructing predictive models.² Their study showed that the obtained models possessed the prowess that can determine the correlation between the lattice constants of the apatite and the ionic radii of given constituent ions. Another application of the ANN³ is the use of multilayer perceptrons with the Levenberg-Marquardt algorithm used for predicting the wear properties of a nickel free stainless steel-HAp. Furthermore, Baseri et al,⁴ investigated variants of multilayer perceptron with varying output units which were used for predicting the mechanical strength of HAp bone cement within specified time durations. Their resulting predictive models showed good performance compared with experimental target values. Previous research highlighted the usage of Sol-gel route⁵ for preparation of HAp and then trained variants of gene expression programming (GEP) approach for predicting HAp crystallite size, given instances of the multivariate input variables; phosphorus precursor, pH level, drying temperature, calcination temperature, aging time, and calcination time. Their results indicated that the GEP-3 yielded the best performance, minimal error and outperformed other approaches. A more comprehensive study was carried out in the research article by Coskun and Karahan.¹ They modeled the corrosion in a bioactive HAp. Their research work compared ANN, GEP, response surface methodology, compared with experimental data. Their result showed that AI models and the experimental data were in close agreement. Moreover, their results findings suggest that the ANN yielded the best-generalized performance compared with other approaches.

Recent research works, which extensively studied the use of deep learning architectures for predicting biomaterials are discussed herein; Vecesi et al⁶ investigated dense and nondense deep learning architecture for executing classification of X-ray diffraction (XRD) patterns from inorganic nonmagnetic materials concerning their crystal system and space group. Their study considered the introduction of noise and created a background signal during data preprocessing. An interesting research article by Oviedo et al⁷ carried out a comprehensive comparison between deep learning and classical supervised learning methods for predicting XRD crystallographic dimensionality and space group and factored data-augmentation in the used spectra data. A slight deviation in the exploration of deep learning was investigated by Hartnett et al;⁸ they studied the unification of data-driven machine learning and density functional theory for predicting formable or thermodynamic stability of iodine which contains apatites.

Other recent research efforts have been directed towards understanding the underlying structure of composite materials or production of materials containing HAp. In the research article by Sharma et al⁹ investigated how organic and inorganic components influence the properties of Bis-GMA/TEGDMA dental composite materials (reinforced from silica nanoparticles and HAp fiber). Another research work by Pande et al¹⁰ presented an approach that produces new composite films based on microstructured HAp particles and polylactic acid. Their research findings suggest that the degradation in temperature conditions of the composite film does not necessarily influence the presence of HAp crystal. An extensive study described different sources of HAp, their respective properties (such as geometry, morphology, particle size, stoichiometry, and thermal stability), and environmental impact on the mode of material synthesis and production.¹¹ In recent times, the dynamic and mechanical assessment of materials have also been studied; for example, Nedaipour et al¹² examined the composition of HAp, polyethylene, glycol and a melt blend of polylactic acid in interference screws; their study analyzed the mechanical and tensile properties of the derived composite relative to material containing pure polylactic acid; the result findings suggest that the derived composite material presents better mechanical properties. Wibisono et al¹³ reported the kinetic model for crab shell milling parameters through the use of ball mills. The model tends to reduce the size of the crab shell before HAp conversion. Their result outcomes highlight discrepancies between the experimental data and their proposed model.

Contribution: The validation of material engineering properties using AI models is gradually becoming a trending area of research and requires less computational complexity compared with traditional approach that requires tedious human engineering. Furthermore, the motivation for the choice of ensemble learning techniques is based on a general opinion that most ensemble learning systems have proven to be effective and extremely versatile in a wide range of research domains.¹⁴ Most of the previous works have employed ANNs and computational intelligent tools for modeling engineering properties of biomaterials. To the best of our knowledge, this is the first time a comparative study will be done using ensemble learning techniques compared with GLM for predicting RI in raw and HAp samples. Based on these

motivations, this article investigate and compared variants of ensemble learning techniques with generalized linear models (GLMs) for predicting the RI of HAp, given Bragg angles as input to the supervised learning methods. The data used are different collections of univariate pair of reflection intensity vs Bragg angles of either raw or HAp samples under varying sintering temperatures. For the experiments, the ensemble techniques based on LightGBM, CatBoost, two variants of XGBoost (tuned with the manual approach and genetic algorithm) were compared with GLM. We carried out 200 experiments by training five methods on each of the four main data lexicons (each data is partitioned into training-testing sets based on 10-fold cross-validation). The results show that all the ensemble learning techniques significantly outperform the GLM on all the examined data. In addition, under varying sintering temperature, variants of the XGBoost yielded best generalization, and lowest error metrics, while outperforming all other approaches. In some data instance, the other ensemble learning methods (CatBoost and LightGBM) experienced difficulty to predict some RI of the HAp; this could have arose due to outlier prediction problem. The GLM models yielded degrading performances and indicate difficulty for the models to understand the underlying patterns in the HAp variations (most especially under varying sintering temperature). It is important to note that most of the ensemble learning predictive outcomes were in close agreement with experimental target output (reflection intensity). Each of the methods computational time was less than 3 minutes as against laboratory generated spectra data which includes interfaces with X-ray generating/radiating equipment which could be tedious (span for hours) and dangerous to human health.

2 | DATASET

In this section, we describe the procedures for the data acquisition and the data distribution employed for carrying out the experiments.

2.1 | Data acquisition

The HAp precursors was derived from nonseparated biowastes (animal bones) which were scavenged from an abattoir dumping site in Zaria, Nigeria. The as-received bones were washed and deprotenized before drying in a hot air oven at 150°C for 8 hours. Next, the powders were grinded and sieved through a mesh size of 300 micrometer. The processed powder was subjected to varying sintering temperatures: 900°C, 1000°C, and 1100°C. Hence the resulting data yielded three variants of the XRD samples; a detailed description of the data and their respective mechanical properties are explained by Obada et al.^{15,16} The illustration of the synthesis of HAp and crystallographic evaluation protocol is shown in Figure 1.

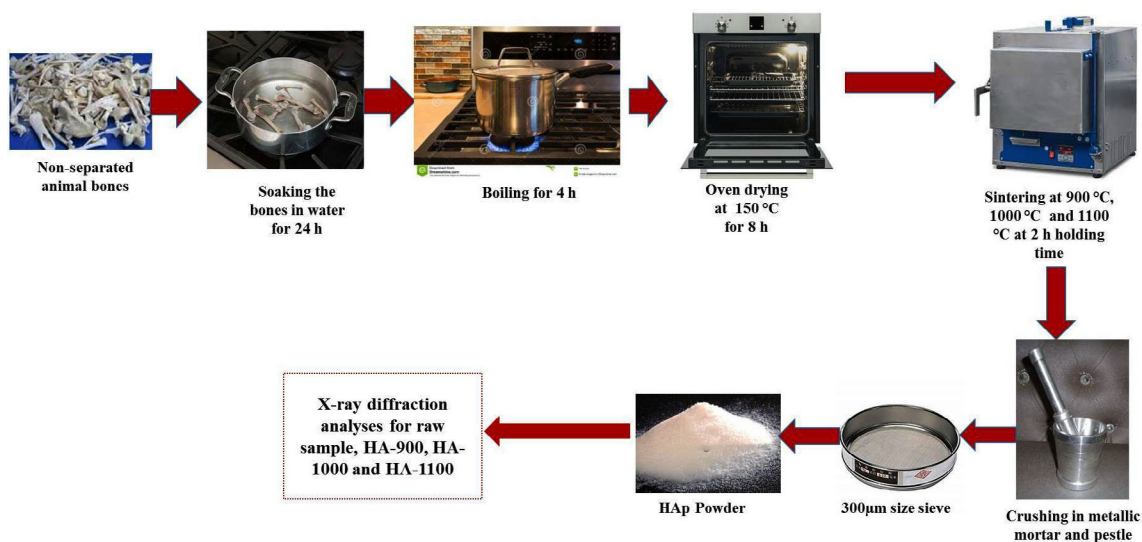


FIGURE 1 HAp processing and crystallographic evaluation protocol.¹⁶ HAp, hydroxyapatite

The resulting data obtained from the XRD experiments for the HAp variants subjected to varying sintering temperatures and raw samples (not subjected to heat treatment) are shown in Figure 2. From the described figures, the boundary condition for the experiments can be drawn; (1) all the X-ray reflections intensities are mapped to the angular bounds (Bragg Angles) in the range $[5^\circ, 90^\circ]$. (2) The reflection intensity levels for the HAp variants depends on the sintering temperature which tailors the crystallinities of the HAp powders.^{15,16} For instance, XRD analysis for the raw sample (no heat treatment) yields lower reflection intensity, but when subjected to sintering temperature with step-wise increment, yields higher RI. This means that an increase in the sintering temperature of HAp under thermal treatment can alter the crystallinities of the materials. To investigate the crystallinity of the synthesized HAp powders, diffraction patterns were collected on a Rigaku Miniflex diffractometer operating a copper tube ($\lambda = 1.5418 \text{ \AA}$) ascertained the crystalline nature of the materials (raw and HAp samples) while providing information on the materials' unit cell dimensions. The goniometer was programmed at a scan rate of $0.033^\circ/\text{s}$ over a 2θ range of 5° to 90° . The materials' crystallinity according to Dey et al¹⁷ can be estimated using the expression as denoted in Equation (1).

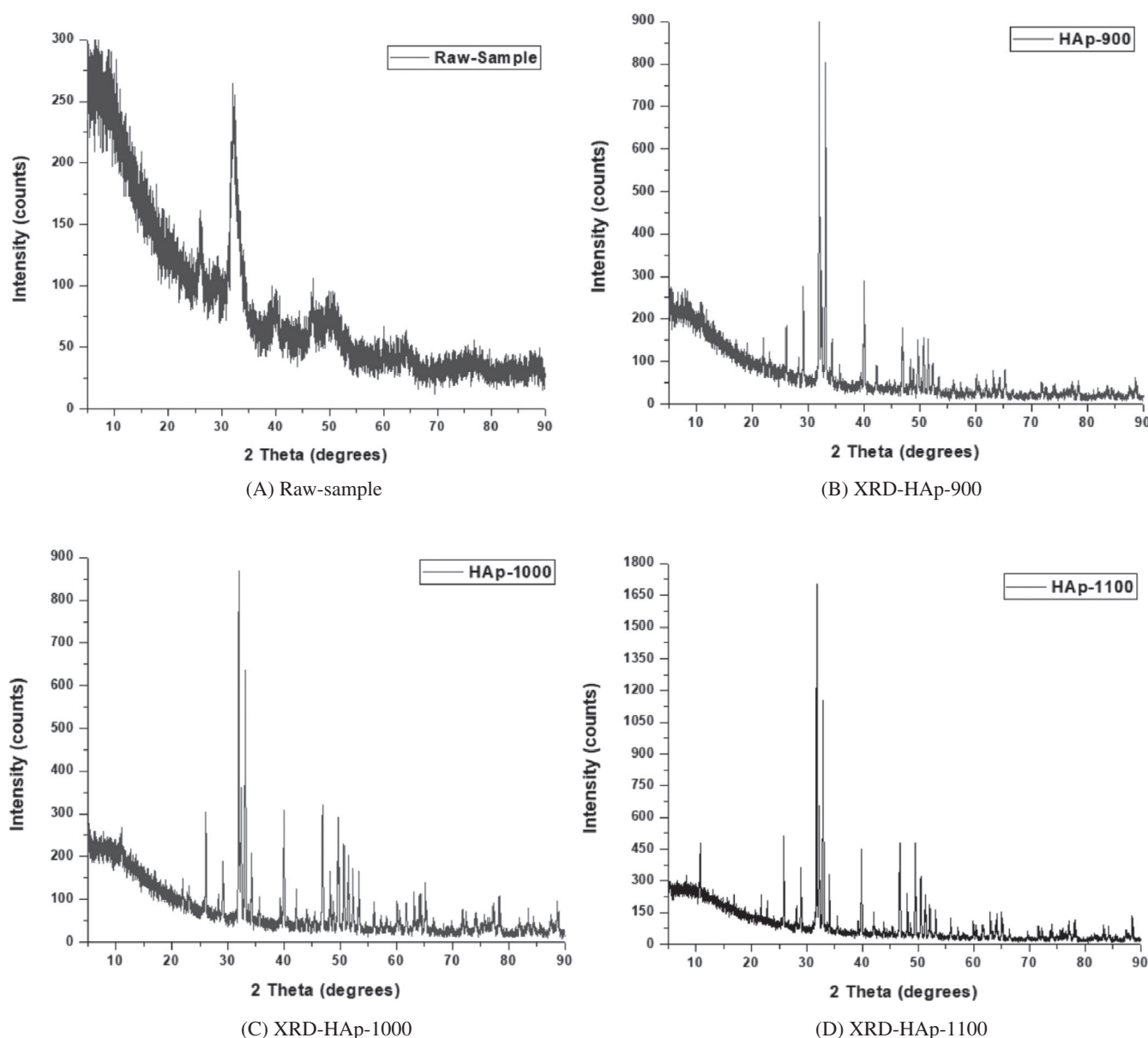


FIGURE 2 Plots of reflection intensity (counts) vs Bragg angles (2θ in degrees) for raw-sample without heat treatment and HAp at different sintering temperature conditions; each of the data lexicons contains 6473 examples as illustrated on a line distribution. HAp, hydroxyapatite

TABLE 1 Cell parameters of the raw sample and HAp

Calcination temperature (°C)	Sample code	Crystallinity	c/a ratio
–	Raw sample	86.65	0.71
900	HAp-900	86.15	0.52
1000	HAp-1000	89.14	0.53
1100	HAp-1100	75.06	0.73

Abbreviation: HAp, hydroxyapatite.

$$X_c = 100 \times \frac{I_{300} - V_{112/300}}{I_{300}}, \quad (1)$$

where I_{300} = intensity of (300) diffraction peak, $V_{112/300}$ = reflection of the hollow between (112) and (300) diffraction peaks of the HAp. The lattice parameters were calculated using Equation (2):

$$\frac{1}{d^2} = \frac{4}{3} \left[\frac{h^2 + k^2 + hk}{a^2} \right] + \frac{l^2}{c^2}, \quad (2)$$

where d accounts for the interplanar distance, the variable h, k, l represents the lattice planes, while a and c are the lattice parameters. The crystallinity of the materials and the ideal axial ratio (c/a) for the materials crystal structure are shown in Table 1.

It is pertinent to note that the HAp samples used in this study were characterized by a rod-like morphology (not shown) with a sixfold symmetry along the c -axis and has a hexagonal shape representation. Moreover, the XRD data showed that the materials are phase pure indicating that the HAp phase is dominant.

2.2 | Dataset distribution

The used data are univariate collection of several pair of input-output distributions of the XRD signatures of either raw or synthesized HAp under varying sintering temperatures. The input accounts for the Bragg angle distribution of the raw or HAp samples, while the target output accounts for the RI in the raw or HAp data. Overall, the four main data lexicons based on the variations of sintering temperature conditions or not were investigated. Each of the data distributions contains 6473 examples and was partitioned into the ratio 80%:20% for training and testing sets, respectively, based on a random shuffle for 10 different folds (10-fold cross-validation), while ensuring no overlapping examples exist in the training and testing set distributions. These distributions imply that a total of 40 different distributions without overlapping were examined for the four data lexicons. This means that based on the five methods examined, a total of 200 experiments were investigated on the four data lexicons. In the current study, we attempt to investigate the use of AI-based supervised learning algorithms to regressively predict the RI given different Bragg angular information of the raw or HAp samples passed as input to the supervised learning methods. The supervised learning methods examined are described in the following section.

3 | SUPERVISED LEARNING TECHNIQUES

In this section, we compared generalized linear model with ensemble learning techniques.

3.1 | Generalized linear model

The GLM¹⁸ is often used for modeling count or binary data. A formulation of the GLM can be described as;

$$y_i = \sum_{k=1}^K \beta_k x_{ki} + \beta_0, \quad (3)$$

where the response y_i is modeled by a linear function of the explanatory variable (input) x_{ki} with input coefficients β_k and bias value β_0 . An invertible link function $g(\cdot)$ operates on the expectation of the dependent variable $\mu_i = E(y_i)$ to a linear predictor (y_i).

$$g(\mu_i) = y_i = \sum_{k=1}^K \beta_k x_{ki} + \beta_0. \quad (4)$$

Since the link function (canonical link) can be expressed in an invertible representation, we can define the mean function as;

$$\mu_i = g^{-1} \left(\sum_{k=1}^K \beta_k x_{ki} + \beta_0 \right). \quad (5)$$

GLM can be assumed as a nonlinear regression model for the dependent variable (response)^a. GLM with a Gamma (Γ) distribution will be used to construct a predictive model. Due to the nature of the CRI distribution in the used data lexicon, can be described as nonlinear. This technique will be considered as the baseline for this study. The GLM conditional variance function $V(\cdot)$ can be computed as $V(y_i/x_{ki}) = \phi \mu_i^2$. The ϕ denotes dispersion parameter.

3.2 | Ensemble methods

3.2.1 | CatBoost

CatBoost is a variant of a gradient boosting technique that employs decision trees as root predictors and it consists of two boosting modes; plain and ordered. The plain-mode is the conventional gradient boosting decision tree (GBDT) algorithm¹⁹ which has an inbuilt ordered target statistics. The ordered-mode is an effective modification of the ordered-boosting algorithm as reported by Prokhorenkova et al.²⁰ The practical implementation of the CatBoost involves the generation of $s + 1$ independent random permutations of examples in a given training data. The permutation value $\sigma_1, \dots, \sigma_s$ are employed when evaluating splits within the defined tree structures. Note that this technique uses a similar partition criterion across every level of the derived trees (decision trees).^{21,22} The derived tree from this technique exhibits the following properties; often balanced, less prone to overfitting, and computationally inexpensive. The ordered boosting mode maintains the supporting model: $M_{r,j}(i)$ is the current prediction for i th example for a given instance of permutation σ_r . For each iteration t of this algorithm, we sample a random permutation σ_r in the bound $\{\sigma_1, \dots, \sigma_s\}$, to construct a tree T_t . A comprehensive mathematical explanation of this method was described in the research article by Prokhorenkova et al.²⁰ The central goal of the CatBoost was implemented to combat prediction shift problems. In the CatBoost experiment, the following parameters were considered; the iterations was initialized to 50 000, maximum depth was set to 3, *learning_rate* = 0.1, the loss function was set to “RMSE,” and other default settings were kept constant.

3.2.2 | LightGBM

LightGBM is another variant of the GBDTs.¹⁹ It employs two components: gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB) for the estimation of information gain using minimal instances of data for a defined feature bundled scenario. The central objective in the design of LightGBM²³ was to introduce a sampling scheme into a GBDT while presenting a good trade-off in data reduction balance and keeping an effective accuracy from the learned decision trees. The earlier mentioned components of the LightGBM play important roles in the feature engineering activity: GOSS execute random sampling of data instances b with small gradients and retains instances a with large gradients. The GOSS computes the information gain by amplifying sampled data with small gradients, which can be defined as $\frac{1-a}{b}$. The EFB employs two algorithms for processing the data features; greedy bundling (GB) and merges exclusive features (MEF). The former (GB) constructs a graph based on weighted edges, then sorts the data features in descending order sequence, and

^ahttps://www.sagepub.com/sites/default/files/upm-binaries/21121_Chapter_15.pdf

inspect the feature order to determine the assignment of features to either existing or new bundles. The latter component (MEF) involves integrating bundle features and ensures a good trade-off in computational complexity (reduction in computing time). A detailed explanation of the described algorithms and some theoretical background for understanding the LightGBM can be found in the research article by Ke et al.²³ In our preliminary experiments, we explored other forms of gradient boosting techniques. The choice of GBDT stems from the observation that the technique yields better performance compared with GOSS, DART, and random forest; moreover, the choosed decision tree present a good performance and tradeoff in computational complexity. The LightGBM system parameters includes; *boosting_type* initialized to “gbdt,” *learning_rate* set to 0.1, *colsample_bytree* set to 1, *max_depth* initialized to -1, *min_child_samples* set to 20, *min_child_weight* initialized to 0.001, *n_estimators* initialized to 50 000, *num_leaves* set to 31, the parameters; *reg_alpha*, *reg_lambda*, *min_split_gain* were initialized to 0.0, *silent* initialized as “True,” and the *subsample* was initialized to 1.0, and all other parameters were set to “None.”

3.2.3 | XGBoost

The XGBoost²⁴ is another variant of gradient boosting method which factors cache access pattern, sharding, and data-compression for building a more scalable ensemble technique (tree boosting system). This technique uses sparsity awareness and weighted quantile sketch algorithms. The latter algorithm approximates the tree learning and the former handles sparse data features. This method process the input features and then constructs a model that can perform any of the following tasks; classification, regression and ranking. We employed the XGBoost for regressive prediction task. A compact mathematical formulation of the loss function^b using the XGBoost hyperparameters can be defined as;

$$L = \sum_{j=1}^{T_m} [G_{jm}w_{jm} + \frac{1}{2}(H_{jm} + \lambda_R)w_{jm}^2 + \alpha_R|w_{jm}|] + \gamma T_m, \quad (6)$$

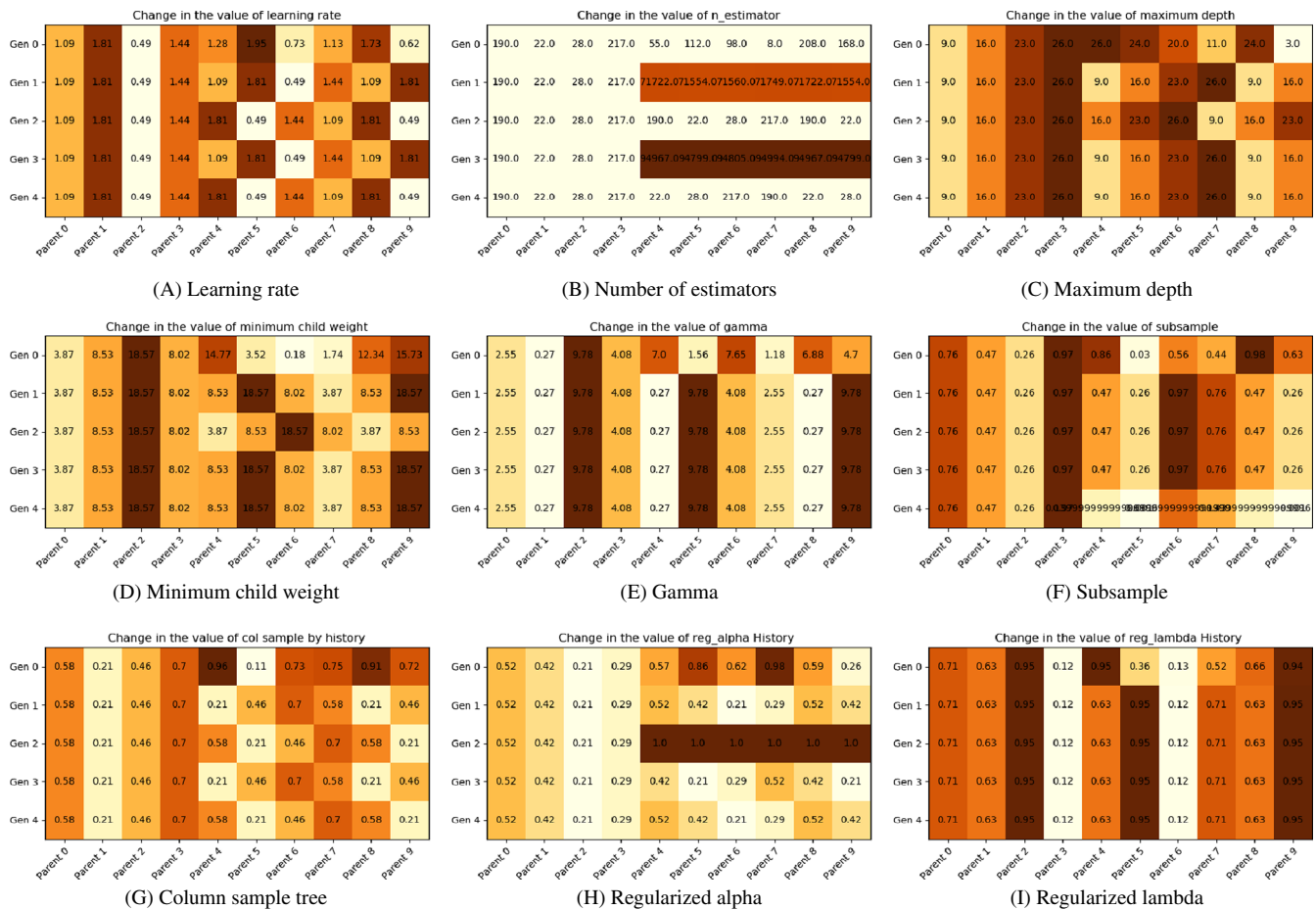
where L is the loss function, α_R and λ_R denote L_1 and L_2 regularization term, respectively, γ denote the penalization term also known as Lagrangian multiplier, it is mainly used for controlling complexity during splitting. The H_{jm} and G_{jm} denote sum of hessian and sum of gradient, respectively, 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 $D_{\max}$. Further explanation of the *colsample* by tree level (S_{col}), *subsample* (S_{sub}), sparse-aware, and shrinkage information were discussed in article by Chen and Guestrin.²⁴ In the experiment, we considered two forms of XGBoost with respect to the tuning of the hyperparameters;

1. Manual tuning; involves the tuning in the number of estimators N_{est} , we searched a nonuniform number of estimator space; {2500, 5000, 7500, ... , 60 000, 200 000} while keeping other hyperparameter constant. The best found number of estimator is 10 000 due to the good performances in both testing and training phases. Hence the resulting method is called XGBoost-Manual Tuning (XGBoost-MT).
2. Genetic Algorithm; GA is a heuristic optimization algorithm that optimizes a variable(s) in a given fitness function using the principle of natural selection and genetic operators (crossover and mutation). GA consists of five main phases;
 - Initialization of population: this phase involves the collection of individual sets, also known as population; each individual denote a solution to a problem. An individual is described by a set of parameters known as genes. Genes merge strings together to form a chromosome (solution). The parameters to be optimized are defined within upper and lower bounds as summarized in Table 2.
 - Fitness function: the second phase helps to determine how fit can an individual compete with other individuals, based on some prediction criteria. The fitness score is often used when training a GA model on train data. In the experiment, we used the F1-score function for predicting fitness individuals.
 - Selection: The third phase entails selecting the fittest individuals and allows the transfer of genes from one generation to the next. Parents (fit individuals) with high fitness scores are selected for mating.

^b<https://towardsdatascience.com/boosting-algorithm-xgboost-4d9ec0207d>

TABLE 2 XGBoost hyperparameters describing upper and lower limits

Hyperparameters bounds	L_r	N_{est}	D_{max}	H_{jm}	γ	S_{sub}	S_{col}	α_R	λ_R
Lower limit	0.001	1	1	0	0.001	0.001	0.001	0.1	0.1
Upper limit	2	100 000	30	20	10	1	1	1	1

**FIGURE 3** Genetic algorithm optimization of XGBoost hyperparameters, where each of the subcaptions from A to I, represent the hyperparameters, each of the first column provides the optimal value in each of the hyperparameters

- Crossover: The fourth phase requires parents for mating, a crossover-point is chosen at random from within the genes of the parents. This process yields offsprings. In the experiment, we considered a uniform crossover due to its simplicity
- Mutation: The last phase involves a scenario where certain genes in an offspring undergo mutation based on low random probability. The essence of mutation is to guarantee diversity in the population pool and prevent poor convergence.

Note that, we constructed heat-map to visualize the GA optimization for each of the hyperparameters fitness change for a specified generations as shown in Figure 3. The summary of the manual and optimized GA hyperparameters of the XGBoost are reported in Table 3.

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 consist of the following hardware specifications; 8 GB of RAM, 2.0 GB of NVIDIA GPU, and CPU (CORE i7). We used an Origin-8 graphical software for plotting the actual and predicted values from the supervised learning models.

TABLE 3 XGBoost hyperparameter tuning

Tuning methods	L_r	N_{est}	D_{max}	H_{jm}	γ	S_{sub}	S_{col}	α_R	λ_R
Genetic algorithm	1.09	190	9.0	3.87	2.55	0.76	0.58	0.52	0.71
Manual tuning	0.07	10 000	3.0	1.5	0	0.6	0.4	0.75	0.45

4 | RESULTS

This section entails the performance metrics and the result analysis for each of the machine learning methods.

4.1 | Performance metrics

The performance metrics that were used to analyze each of the machine learning techniques are detailed below;

1. Coefficient of determination: The coefficient of determination popularly referred to as R^2 is a statistical tool used for finding the degree of variability between two sets of variables^c. The R^2 is also known as the goodness of fit. The value of R^2 lies between $\{0, 1\}$. A model is said to be a perfect fit and highly reliable if $R^2 = 1$. By contrast, if a model depicts a failure to construct a hypothesis that predicts a data, then $R^2 = 0$. The formulation of R^2 can be defined as:

$$R^2 = 1 - \frac{\sum_i (y_i - y_i^p)^2}{\sum_i (y_i)^2}, \quad (7)$$

where y_i denote the target values and y_i^p represents the predictive values from each of the supervised learning algorithms previously described. Additionally each of the examples of the intensity strength in the training or testing phase can be denoted as i .

2. Root mean square error (RMSE): The RMSE^d computes the differences between a hypothetical predicted model outcome and actual experimental values. RMSE often determines the goodness of fit from a generalized regression model. The RMSE can be defined as;

$$RMSE = \left(\frac{1}{n} \sum_i (y_i - y_i^p)^2 \right)^{\frac{1}{2}}. \quad (8)$$

3. Mean absolute error (MAE): MAE is a statistical metric for computing the absolute difference between two continuous variables. In other words, it is a measure of the difference between the predicted model and the actual experimental data. MAE can be defined as;

$$MAE = \frac{1}{n} \sum_i (y_i - y_i^p). \quad (9)$$

4.2 | Results analysis and discussion

4.2.1 | Result analysis

All the plots shown in Figures 4A-D and 5A-D were based on $\frac{1}{10}$ th of the total experimental runs in the testing phase. After 10 experimental runs (10-fold cross-validations) for each of the methods per data lexicon, we computed the effective average plus the SDs for the performance metrics (R^2 , MAE, RMSE) in both training and testing phases were reported in Tables 4 to 7.

^c<https://www.investopedia.com/terms/c/coefficient-of-determination.asp>

^d<https://methods.sagepub.com/reference/encyc-of-research-design/n392.xml>

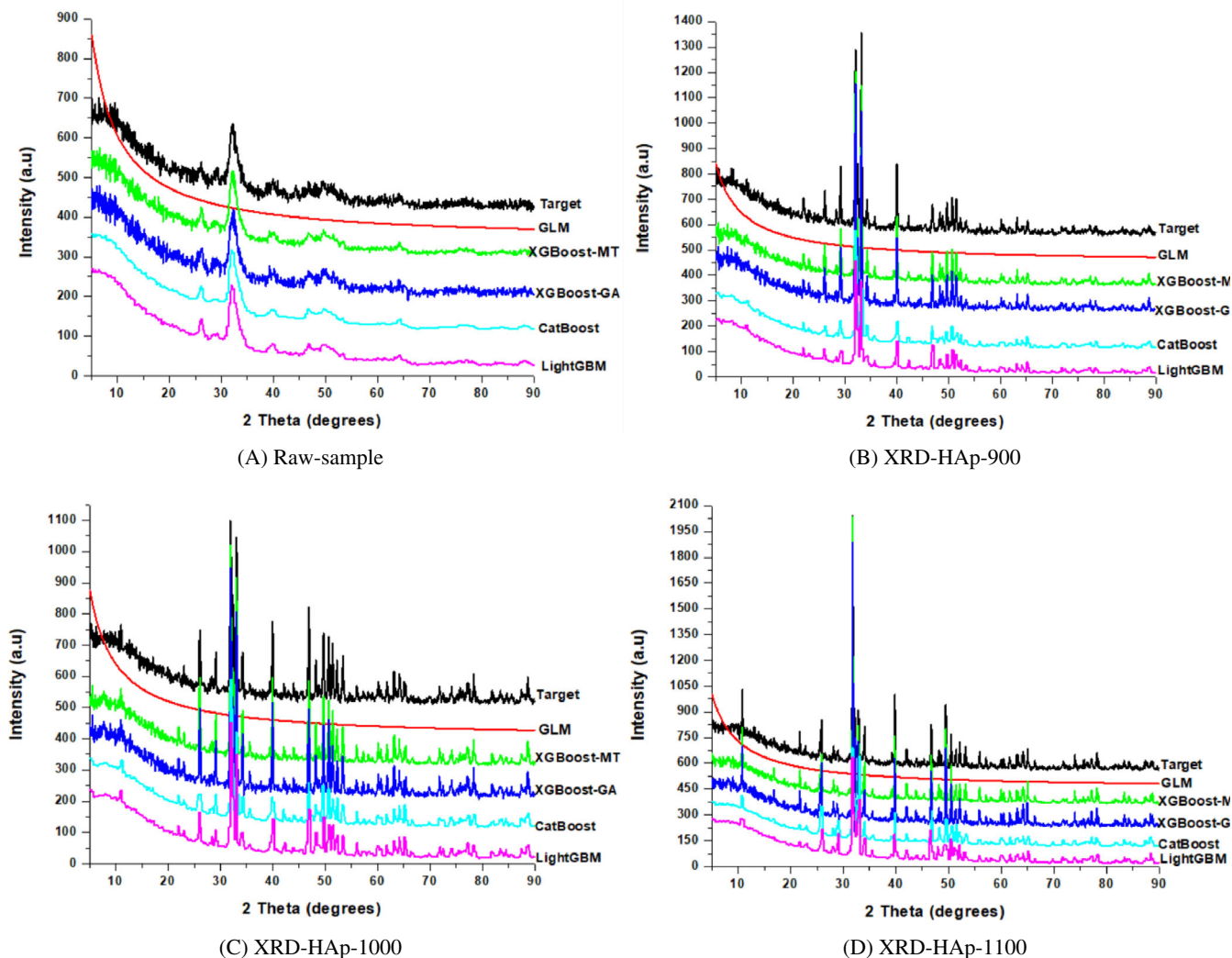


FIGURE 4 Testing phase performance visualization for each of the methods, showing the plots 1/10 experiments (sixth run) of reflection intensity vs Bragg angles based on the raw data sample and synthesized HAp under different sintering temperature conditions (900°C, 1000°C, and 1100°C) on each testing set that contains 1295 examples. HAp, hydroxyapatite

The details of our research findings are highlighted below;

1. All the plots shown in Figure 4 illustrate the visual representation of the ensemble learning techniques and GLM models. An observation can be drawn; the variants of the XGBoost displayed excellent prediction and overlap on the actual target (reflection intensity) in almost all the plots, while Figure 4B-D indicate that other ensemble learning methods (CatBoost and LightGBM) show good prediction on modal RI (at low peak) but experienced difficulty in prediction on high peak RI in the testing phase. The aforementioned problem may have arisen due to outlier prediction problem. All the ensemble learning methods significantly outperform GLM at $P < .05$, and we remark all the ensemble models were reliable and depict a high goodness fit. The GLM in all the plots showed a Zipf's law curve but with the lowest reliability in predicting the RI when compared with the ensemble learning approaches.
2. Hence to further measure the model goodness fit for each of the methods, we visualized the model correlation in Figure 5. The plots in this figure show the correlation plots for each of the methods. From Figure 5A shows the correlation plots on the raw data sample; here we report that all the ensemble learning methods yielded a well-correlated performance ($R^2 \geq 0.96$) and a predictive error {MAE, RMSE} < 13.2 when compared with GLM model whose ($R^2 \leq 0.667$) and a higher predictive error MAE, RMSE ≥ 18.04 in both the training and testing sets, respectively. The performance of all the ensemble learning methods on the raw sample is said to be very reliable and demonstrated

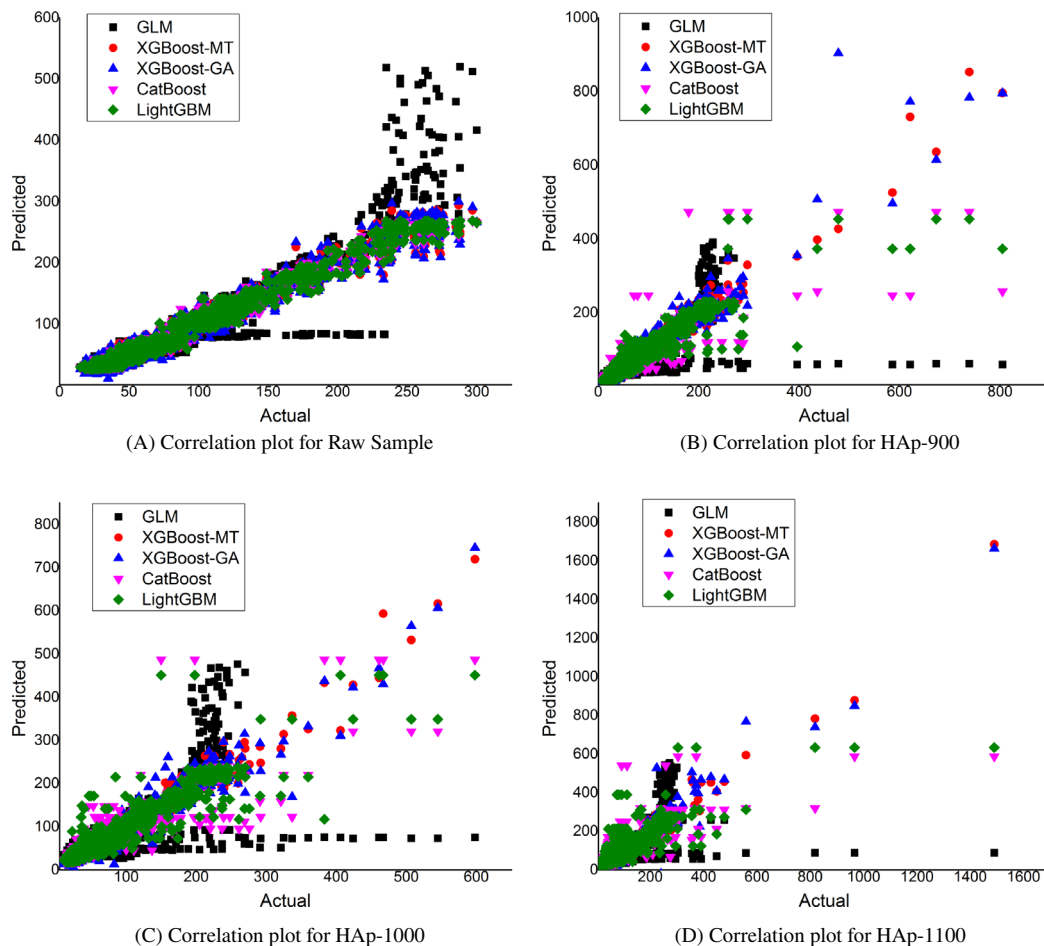


FIGURE 5 Testing phase correlation plot for each of the methods in 1/10 experiments (sixth run) on each of the main data lexicons, while showing the correlation relationship between actual target output and predicted outputs from the supervised learning models

a good fit; while, the lower performance of the GLM can be ascribed due to outlier prediction (false-prediction). The summary of the described information are presented in Table 4.

- From Figure 5B on a HAp-900 data sample, the results show that the variants of the XGBoost yielded a high correlation score: $R^2 \geq 0.94$ and a minimal predictive error $MAE \leq 9.19$ and $RMSE \leq 15.85$ compared with other ensemble learning approaches (LightGBM and CatBoost) which presents; higher predictive error $MAE \leq 11.2$ and $RMSE \leq 29.17$ while indicating above-average correlation score $R^2 \leq 0.838$ in both testing and training phases. Similarly as in the previous explanation, all the ensemble learning methods demonstrated reliable fitness scores compared with the GLM which yielded an average correlation score with higher predictive errors in both evaluation phases. The summary of the described information are presented in Table 5.
- From Figure 5C on a HAp-1000 data sample, the results show that the variants of the XGBoost yielded a high correlation score: $R^2 \geq 0.94$ and a minimal predictive error $MAE \leq 10.18$ and $RMSE \leq 17.35$ compared with other ensemble learning approaches (LightGBM and CatBoost) which presents; higher predictive error $MAE \leq 13.3$ and $RMSE \leq 30.04$ while indicating above-average correlation score $R^2 \leq 0.836$ in both testing and training phases. Again, all the ensemble learning methods demonstrated reliable fitness scores compared with the GLM which yielded a below-average correlation score with a more higher predictive errors in both evaluation phases. The summary of the described information are presented in Table 6.
- From Figure 5D on a HAp-1100 data sample, the results show that the variants of the XGBoost yielded a high correlation score: $R^2 \geq 0.93$ and a minimal predictive error $MAE \leq 12.55$ and $RMSE \leq 26.53$ compared with other ensemble learning approaches (LightGBM and CatBoost) which presents; a much higher predictive error $MAE \leq 19.24$ and $RMSE \leq 53.83$ while indicating above-average correlation score $R^2 \leq 0.728$ in testing phases. Again, all the ensemble learning methods demonstrated reliable fitness scores compared with the GLM which obtained the worst

TABLE 4 Performance evaluation metric comparison of the different techniques on Raw-sample data after 10-folds cross-validation, showing the mean and SD of the metrics

Techniques	Train			Test		
	MAE	RMSE	R^2	MAE	RMSE	R^2
XGBoost-GA	1.702 ± 0.058	2.245 ± 0.092	0.999 ± 0.000	9.677 ± 0.072	13.131 ± 0.096	0.961 ± 0.001
XGBoost-MT	4.149 ± 0.077	5.220 ± 0.099	0.994 ± 0.000	8.241 ± 0.121	11.369 ± 0.179	0.971 ± 0.001
LightGBM	7.228 ± 0.038	9.761 ± 0.042	0.978 ± 0.000	7.432 ± 0.141	10.024 ± 0.252	0.977 ± 0.001
CatBoost	7.308 ± 0.067	9.840 ± 0.078	0.978 ± 0.000	7.441 ± 0.098	10.022 ± 0.187	0.977 ± 0.001
GLM	18.041 ± 0.184	38.349 ± 0.489	0.667 ± 0.009	18.237 ± 0.604	38.749 ± 1.808	0.660 ± 0.034

Abbreviations: GLM, generalized linear model; MAE, mean absolute error; RMSE, root mean square error.

TABLE 5 Performance evaluation metric comparison of the different techniques on XRD-900-HAp data after 10-folds cv

Techniques	Train			Test		
	MAE	RMSE	R^2	MAE	RMSE	R^2
XGBoost-GA	1.435 ± 0.058	1.874 ± 0.088	0.999 ± 0.000	9.191 ± 0.272	15.849 ± 2.199	0.947 ± 0.016
XGBoost-MT	3.153 ± 0.025	3.975 ± 0.030	0.997 ± 0.000	7.490 ± 0.212	11.263 ± 0.517	0.974 ± 0.002
CatBoost	10.814 ± 0.558	28.531 ± 1.441	0.838 ± 0.016	11.201 ± 1.202	27.764 ± 4.667	0.838 ± 0.045
LightGBM	10.558 ± 0.364	28.658 ± 1.819	0.836 ± 0.019	11.062 ± 0.790	29.170 ± 4.374	0.821 ± 0.050
GLM	18.667 ± 0.347	49.807 ± 0.995	0.506 ± 0.011	18.785 ± 0.838	48.263 ± 3.636	0.519 ± 0.037

Abbreviations: cv, cross-validation; GLM, generalized linear model; HAp, hydroxyapatite; MAE, mean absolute error; RMSE, root mean square error.

TABLE 6 Performance evaluation metric comparison of the different techniques on XRD-1000-HAp data after 10-folds cv

Techniques	Train			Test		
	MAE	RMSE	R^2	MAE	RMSE	R^2
XGBoost-GA	1.651 ± 0.069	2.179 ± 0.095	0.999 ± 0.000	10.177 ± 0.292	17.346 ± 2.124	0.940 ± 0.014
XGBoost-MT	3.471 ± 0.074	4.405 ± 0.104	0.996 ± 0.000	8.185 ± 0.168	11.925 ± 0.454	0.972 ± 0.002
CatBoost	12.679 ± 0.412	28.541 ± 1.665	0.839 ± 0.019	13.216 ± 0.655	28.823 ± 1.374	0.836 ± 0.016
LightGBM	12.691 ± 0.288	28.823 ± 1.375	0.836 ± 0.016	13.302 ± 0.496	30.043 ± 3.556	0.819 ± 0.036
GLM	27.105 ± 0.376	56.614 ± 0.811	0.369 ± 0.009	27.529 ± 0.827	56.853 ± 2.906	0.357 ± 0.029

Abbreviations: cv, cross-validation; GLM, generalized linear model; HAp, hydroxyapatite; MAE, mean absolute error; RMSE, root mean square error.

correlation score with highest predictive errors in both evaluation phases. The summary of the described information are presented in Table 7.

4.2.2 | Result discussion

The core objective of the study is to determine which method presents the best predictive model (highest linear correlation: diagonal scatter plot) from the examined supervised learning methods when predicting the reflection intensity of HAp or raw data samples, and a corresponding minimum predictive error. As summarized in Tables 4 to 7, from the results reported in all the tables, the XGBoost (when tuned manually or using heuristic method) yielded performance that surpass all other techniques in the training phase on all the used data; while showing that the XGBoost presented the best correlated model which approximately attained a mean value of $R^2 \sim 1$. A clear observation can be drawn from the results reported in Tables 4 to 7, in the testing phase, out of 180/200 experiments, XGBoost-MT yielded a mean value of $R^2 \geq 0.971$ with relatively low predictive errors on most of used data samples; most especially for sample data subjected to varying sintering temperature conditions. From Tables 5 to 7); all the XGBoost-GA also yielded a highly correlated score; $R^2 \geq 0.93$, which indicates that the model shows an excellent generalization capacity. Overall, we report that XGBoost

TABLE 7 Performance evaluation metric comparison of the different techniques on XRD-1100-HAp data after 10-folds cv

Techniques	Train			Test		
	MAE	RMSE	R^2	MAE	RMSE	R^2
XGBoost-GA	1.889 ± 0.072	2.517 ± 0.108	0.999 ± 0.000	12.551 ± 0.708	26.531 ± 4.930	0.933 ± 0.015
XGBoost-MT	4.054 ± 0.125	5.199 ± 0.175	0.997 ± 0.000	9.460 ± 0.323	15.501 ± 1.291	0.977 ± 0.002
LightGBM	17.042 ± 0.696	47.738 ± 3.021	0.750 ± 0.019	18.792 ± 1.696	53.827 ± 10.117	0.728 ± 0.051
CatBoost	17.768 ± 2.400	48.541 ± 4.702	0.741 ± 0.037	19.238 ± 1.796	54.292 ± 8.095	0.723 ± 0.037
GLM	30.826 ± 1.174	76.666 ± 4.229	0.356 ± 0.031	32.569 ± 2.027	85.545 ± 13.806	0.315 ± 0.079

Abbreviations: cv, cross-validation; GLM, generalized linear model; HAp, hydroxyapatite; MAE, mean absolute error; RMSE, root mean square error.

(in both tuning conditions) yielded the best goodness fit in predicting the reflection intensity at relatively low predictive error on almost all the used data. The next competitive techniques are; the LightGBM and CatBoost which marginally outperform the variants of the XGBoost on the raw-sample data as shown in Table 4. This is because the correlation score attained $R^2 = 0.977$ with the least predictive error $MAE = 7.4$ in the testing phase. We remark that the XGBoost-GA models obtained lower performance index when compared with XGBoost-MT in the testing phase relative to their respective training predictive outcomes; the reason can be attributed due to overfitting problem. In addition, the second-best technique is the CatBoost which outperforms LightGBM and GLM on data samples HAp-900 and HAp-1000, but not the case on data sample HAp-1100. Moreover, the results obtained from the GLM yielded degrading performance when compared to all the ensemble techniques on the variants of the used data, while obtaining the worst predictive error metrics (yielded maximum MAE and RMSE errors). The reason for the lowest performance from the GLM could be ascribed due to two reasons; (1) GLM employed only one base estimator method compared with ensemble techniques which utilized multi-estimators with an advantage of majority voting scheme during prediction of the reflection intensity spectra. (2) Another reason for the GLM low performance can be explained by the inference drawn from Figure 5; were most of the instances of the GLM predicted more outliers than other approaches (ensemble learning methods). We observed that the GLM model depicted a Zipf's law curve. We report that, all the supervised learning algorithms examined during prediction of reflection intensity, for each run per method, the computational time is less than 3 minutes; relative to laboratory experimental methods which stems from material synthesis protocol to experimental equipment-driven investigations involve lots of tedious engineering processes (can span for hours).

5 | CONCLUSION

This research has compared ensemble learning techniques with GLM for predicting reflection intensity of natural raw sample and HAp under varying sintering temperature conditions. Given the Bragg angular information and actual target output (continuous labels: reflection intensity) passed as input to the supervised learning algorithms based on the expectation to construct several predictive models. The study demonstrated that in most cases the ensemble learning methods obtained results that are in close agreement with experimental data (normal engineering process), in the context of goodness fit and predictive errors. In addition, the research findings show that the XGBoost model yielded a robust performance that surpasses other approaches during the training phase. Furthermore, in the testing phase, all variants of the XGBoost (with manually and GA tuned) outperform other methods, most especially under sintering temperature conditions. However, this is not the case for the raw sample which was not subjected to heat treatment.

The XGBoost-MT used a greater number of estimators which allowed the model to obtain the highest performance most especially during the testing phase. However, the heuristic inspired method (XGBoost-GA) obtained good performance with a limited number of estimators. Another competitive method is the CatBoost which outperforms both LightGBM and GLM under some instances of the sintering temperature conditions (900°C and 1000°C). Although in some cases, the performance gain in the CatBoost is marginal when compared with LightGBM. Furthermore, the research finding showed that the GLM yielded the worst performance when compared with all the ensemble learning techniques. The GLM performance could be attributed to the use of a limited number of predictive estimators and finds it difficult to understand the underlying patterns within the variants of the used HAp data. All the examined supervised learning algorithms present an advantage of low computational complexity compared with laboratory method which is very tedious and time-consuming.

The overall knowledge gained from this study suggests that the AI model based on ensemble learning can be deployed to the material and biomedical engineering field for building predictive models. The training of XGBoost on variants of the univariate continuous data (HAp and raw samples) demonstrated very good learning of the underlying patterns compared with other ensemble learning methods (LightGBM and CatBoost). However, we do not imply that XGBoost is always better than training other ensemble learning techniques on multivariate data representation. In our study, we report that each of the dataset variants contains less than 6.5K examples on a univariate distribution of the XRD patterns for a predefined range of Bragg angles (2θ). The examined machine learning models were constrained to a regressive prediction of XRD patterns (intensity) on univariate data lexicons (one input feature dimension). More research work can explore the possibility to construct a model that can predict XRD patterns on a multivariate input feature space (containing different material engineering properties). In addition, it will be interesting to use a more scalable data lexicon, to further assess the benefit of ensemble learning, deep learning, and other machine learning techniques. Finally, future research can investigate the training of machine learning algorithms on XRD data at higher degree of sintering temperature conditions ($> 1200^{\circ}\text{C}$).

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.

PEER REVIEW INFORMATION

Engineering Reports thanks Prasanna Balachandran, Sankar Chandra Basu, John Ferrie, and other anonymous reviewer(s) for their contribution to the peer review of this work.

PEER REVIEW

The peer review history for this article is available at <https://publons.com/publon/10.1002/eng2.12292/>.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

AUTHOR CONTRIBUTIONS

Emmanuel Okafor: Conceptualization-Lead; Formal analysis-Lead; Investigation-Lead; Methodology-Lead; Software-Lead; Validation-Lead; Visualization-Lead; Writing-original draft-Lead. **David Obada:** Data curation-Supporting; Methodology-Supporting; Writing-review, editing-Supporting. **Yusuf Ibrahim:** Methodology-Supporting; Writing-review, editing-Supporting. **David Dodoo-Arhin:** Conceptualization-Equal; Data curation-Supporting; Formal analysis-Lead; Methodology-Supporting; Project administration-Equal; Supervision-Equal; Writing-review, editing-Supporting.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Emmanuel Okafor  <https://orcid.org/0000-0001-6929-6880>

David Dodoo-Arhin  <https://orcid.org/0000-0002-9657-2489>

REFERENCES

1. Coşkun Mİ, Karahan İH. Modeling corrosion performance of the hydroxyapatite coated CoCrMo biomaterial alloys. *J Alloys Compound*. 2018;745:840-848.
2. Wu P, Zeng YZ, Wang CM. Prediction of apatite lattice constants from their constituent elemental radii and artificial intelligence methods. *Biomaterials*. 2004;25(6):1123-1130.
3. Younesi M, Bahrololoom ME, Ahmadzadeh M. Prediction of wear behaviors of nickel free stainless steel-hydroxyapatite bio-composites using artificial neural network. *Comput Mater Sci*. 2010;47(3):645-654.
4. Baseri H, Rabiee SM, Moztafzadeh F, Solati-Hashjin M. Mechanical strength and setting times estimation of hydroxyapatite cement by using neural network. *Mater Des (1980-2015)*. 2010;31(5):2585-2591.

5. Jafari MM, Khayati GR. Prediction of hydroxyapatite crystallite size prepared by sol-gel route: gene expression programming approach. *J Sol-Gel Sci Technol*. 2018;86(1):112-125.
6. Vecsei PM, Choo K, Chang J, Neupert T. Neural network based classification of crystal symmetries from x-ray diffraction patterns. *Phys Rev B*. 2019;99(24):245120.
7. Oviedo F, Ren Z, Sun S, et al. Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. *NPJ Comput Mater*. 2019;5(1):1-9.
8. Hartnett TQ, Ayyasamy MV, Balachandran PV. Prediction of new iodine-containing apatites using machine learning and density functional theory. *MRS Commun*. 2019;9(3):882-890.
9. Sharma S, Kumar P, Chandra R, Setia P. Prediction of properties of silica nanoparticle/hydroxyapatite fiber reinforced Bis-GMA/TEGDMA composites using molecular dynamics. *Comput MaterSci*. 2019;158:32-41.
10. Pandele AM, Constantinescu A, Radu IC, Miculescu F, Ioan Voicu S, Ciocan LT. Synthesis and characterization of pla-micro-structured hydroxyapatite composite films. *Materials*. 2020;13(2):274.
11. Agbeboh NI, Oladele IO, Daramola OO, Adediran AA, Olasukanmi OO, Tanimola MO. Environmentally sustainable processes for the synthesis of hydroxyapatite. *Heliyon*. 2020;6(4):e03765.
12. Nedaipour F, Bagheri H, Mohammadi S. "Polylactic acid-polyethylene glycol-hydroxyapatite composite" an efficient composition for interference screws. *Nanocomposites*. 2020;6(3):1-12.
13. Wibisono Y., Rafianto V., Alvianto D., Bagus Hermanto M. Prediction of size reduction by batch ball milling process for crab shell powder prior hydroxyapatite conversion. *Paper presented at: IOP Conference Series: Earth and Environmental Science*. 2020;542:012011. <http://dx.doi.org/10.1088/1755-1315/542/1/012011>.
14. Polikar R. *Ensemble Learning*. In: Zhang C., Ma Y. (eds) *Ensemble Machine Learning*. Boston, MA: Springer; 2012:1-34. https://doi.org/10.1007/978-1-4419-9326-7_1.
15. Abifarin JK, Obada DO, Dauda ET, Dodoo-Arhin D. Experimental data on the characterization of hydroxyapatite synthesized from biowastes. *Data Brief*. 2019;26:104485.
16. Obada DO, Dauda ET, Abifarin JK, Dodoo-Arhin D, Bansod ND. Mechanical properties of natural hydroxyapatite using low cold compaction pressure: effect of sintering temperature. *Mater Chem Phys*. 2020;239:122099.
17. Dey S, Das M, Balla VK. Effect of hydroxyapatite particle size, morphology and crystallinity on proliferation of colon cancer HCT116 cells. *Mater Sci Eng C*. 2014;39:336-339.
18. Nelder JA, Wedderburn RWM. Generalized linear models. *J Royal Stat Soc Ser A (General)*. 1972;135(3):370-384.
19. Friedman J H. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*. 2001;29(5):1189-1232. <http://dx.doi.org/10.1214/aos/1013203451>.
20. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems 31*. Paper presented at: 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada; 2018:6638-6648. <https://papers.nips.cc/paper/7898-catboost-unbiased-boosting-with-categorical-features.pdf>. Accessed August 23, 2020.
21. Ferov M, Modrý M. Enhancing lambdamart using oblivious trees; 2016:1-12. arXiv preprint arXiv:1609.05610. <https://www.semanticscholar.org/paper/Enhancing-LambdaMART-Using-Oblivious-Trees-Ferov-Modr%C3%BD/924e59055a24b5f7c981ca6f8adfec170a2a360b>. Accessed August 23, 2020.
22. Gulin A, Kuralenok I, Pavlov D. Winning the transfer learning track of yahoo!'s learning to rank challenge with yetirank. *Proceedings of the Learning to Rank Challenge, Haifa, Israel; PMLR 14*, 2011:63-76. <http://proceedings.mlr.press/v14/gulin11a.html>. Accessed August 23, 2020.
23. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. LightGBM: a highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems: Long Beach, CA, USA: 31st Conference on Neural Information Processing Systems (NIPS 2017), Neural Information Processing Systems Foundation, Inc; 2017:3146-3154*.
24. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. Paper presented at: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA; 2016:785-794*.

AUTHOR BIOGRAPHIES

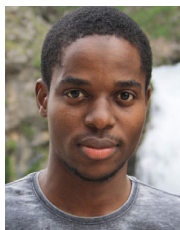


Dr. Emmanuel Okafor is a Lecturer in the field of machine learning, computer vision, and control engineering at the Department of Computer Engineering, Ahmadu Bello University (ABU), Nigeria. He earned Bachelor's Degree in Electrical Engineering (2010) and Master's Degree in Control Engineering (2014) from ABU. Dr. Okafor obtained a Ph.D. in the field of Artificial Intelligence (2019) from the University of Groningen, the Netherlands. His current research interests include: computer vision, deep learning, time series forecasting, reinforcement learning, robotics, and optimization.



Dr. David O. Obada is a Lecturer at the Department of Mechanical Engineering, Ahmadu Bello University, Zaria, Nigeria. He graduated with a B.Tech in mechanical engineering at the Ladoke Akintola University of Technology, Ogbomoso, Nigeria, and obtained M.Sc. and Ph. D degrees at the Ahmadu Bello University, Zaria, Nigeria specializing in production engineering. His research interests include fracture mechanics, advanced materials, and internal combustion engines. In addition to his teaching, research, student's supervision and several refereed publications, Dr. Obada is a beneficiary of internationally recognized awards. Dr. Obada participated in the MIT-Empowering the Teachers Program at the Massachusetts Institute of Technology (MIT),

USA and holds a Kaufmann Teaching Certificate also from MIT.



Dr. Yusuf Ibrahim is at present a Lecturer in the Department of Computer Engineering, Ahmadu Bello University, Zaria, Nigeria. He received his Bachelor's Degree in Electrical Engineering from Ahmadu Bello University, Zaria, Nigeria in 2011. He also obtained his Master's and Doctor's Degrees in Computer Engineering in the years 2017 and 2019, respectively, from the same institution. His research interests include Machine Learning, Digital Image Processing and Computer Vision. Yusuf is a registered member of the Council for the Regulation of Engineering in Nigeria (COREN).



Prof. David Dodoo-Arhin is an Associate Professor in the Department of Materials Science and Engineering, School of Engineering Sciences, University of Ghana. He Holds a BSc degree in Physics and post-graduate Diploma in Science Education (PGDE) from the University of Cape Coast, Ghana. He obtained his PhD degree in Materials Engineering from the Department of Materials Engineering and Industrial Technologies, University of Trento-Italy in 2010 with particular focus on nanostructured oxides for energy and environmental applications. His research areas include materials for energy generation, conversion and storage, nanotechnology and engineering ceramics.

How to cite this article: Okafor E, Obada DO, Ibrahim Y, Dodoo-Arhin D. Prediction of the reflection intensity of natural hydroxyapatite using generalized linear model and ensemble learning methods. *Engineering Reports*. 2021;3:e12292. <https://doi.org/10.1002/eng2.12292>