RESEARCH PAPER

Atomic Coordinate Prediction of Carbon Nanotubes Using Regression Tree Ensembles

Abdulrahman Saleh Ibrahim [1]*, ***Omar Ibrahim Obaid*** [2], ***Ruaa Yousif Hassan*** [3], ***Thamir K. Al-Azawi*** [3]

[1] *Construction and building engineering technology, Al esraa university college, Iraq*

[2] *Department of Computer, College of Education, AL-Iraqia University, Iraq*

[3] *Al esraa university college, Iraq*

ARTICLE INFO

Article History:
Received 13 July 2023
Accepted 19 September 2023
Published 01 October 2023

Keywords:
Atomic Coordinate
Carbon Nanotubes
Prediction
Machine Learning (ML)
Regression Tree Ensembles

ABSTRACT

In this paper, regression tree ensembles (bagged and boosted) have been utilized in predicting atomic coordinate of Carbone nanotubes (CNTs). The aim of this study is to use ensembles classifiers to compute the atomic coordinates of Carbone nanotubes rather than other simulation tools. The dataset we used in this paper are provided by the UCI Repository of Machine Learning and it has a total of (10721) instances with (8) attributes (five as inputs and three as outputs) and it has no missing data. Various performance measures are also calculated to evaluate the classifiers we employed. The results show that there is a slight difference in performance between bagged and boosted trees, however, they are preferable classifiers for carbon atom coordinates prediction due to their high accuracy and short computation time. Using these predicted atomic coordinates as early coordinates for the simulation tool, the actual atomic coordinates can be retrieved in minutes or seconds instead of days by minimizing the iterations in the computation process.

How to cite this article
Ibrahim A., Obaid O., Hassan R.,Al-Azawi T. Atomic Coordinate Prediction of Carbon Nanotubes Using Regression Tree Ensembles. J Nanostruct, 2023; 13(4):1020-1030. DOI: 10.22052/JNS.2023.04.011

INTRODUCTION

To tackle the issues created by miniaturization, carbon nanotubes (CNTs) have been introduced as an alternative to copper/aluminum metallic interconnects. CNTs are rolled-up sheets of 2-D graphene crystal. Since they are rolled up, they have electronic components that vary based on their direction [1,2]. For years, ab initio estimates have had a major effect on the investigation of material properties. Ab initio methods have no parameters and only need the atomic number as input. These are the reasons for ab initio methods' great success. Through advances in computer efficiency and algorithms, these methods are now being applied to a growing number of physical and chemical phenomena [2, 3]. Ab initio computational methods are used to estimate chemical and physical characteristics of periodic systems in terms of chemical composition and crystalline structure as accurately as possible at a good cost without any requirement for a given empirical data [4].

BIOVIA Materials Studio CASTEP is a leading code for computing the characteristics of materials from fundamental principles. Density functional theory can be used to model a wide range of material properties. Structural at the atomic scale, electronic structure, electrical responsiveness,

* *Corresponding Author Email: abdalrahman@esraa.edu.iq*

and vibration characteristics are some of these properties [5, 2]. The computation time for a huge number of atoms is exceedingly long, as shown in numerous CASTEP simulations. Due to the power of the computer/server, computations may take many days to complete. With density functional theory, the fastest and most accurate approach to calculate atomic coordinates, atomic coordinates are calculated faster than with any other mathematical approach. In contrary, users must employ more powerful computers and parallelism machines, both of which are prohibitively expensive to minimize computing time [2].

Rather than computing the problem, we take a different strategy here; the atomic coordinates are precisely predicted in such a little period. The actual atomic coordinates can be obtained in minutes or seconds instead of days using these anticipated atomic coordinates as early coordinates for the simulation tool.

The use of approximated atomic coordinates in research also enhances the efficiency of acquiring appropriate results. In order to acquire rapid and accurate results, we focused our research on machine learning approaches that have been utilized in the literature for forecasting such tasks. To transfer learning researches to machine learning, several paradigms and techniques are deployed. Statistic recognition systems, symbolic processing, case-based learning, artificial neural networks, and evolutionary programming are just a few examples. Machine learning's goal can be characterized as computers performing the task of human learning. Variety of approaches and algorithms are utilized during this learning [6, 2].

Numerous problems can be solved by performing long and tough formulas in an actual or simulation tools, and the outcomes can be obtained within that approach. In most cases, these situations result in lengthy and costly hardware and software development. Instead, artificial neural networks (ANNs) approaches are being utilized to forecast the outcomes of these issues utilizing datasets received from the real world or a simulated environment.

Artificial Neural Network (ANN) is used by Cheng et al. [7] to construct a model on a graphene metal–oxide–semiconductor field effect transistor. The computing time for the MOSFET model was considerably reduced. The graphene MOSFET model was implemented as a subcircuit in HSPICE software, which may obviously improve the efficiency of simulations on graphene large-scale incorporated circuits. Yet, another work by Cheng et al. [8] has coupled support vector regression (SVR) along with particle swarm optimization to build mathematical models to predict mechanical characteristics of carbon nanotubes/epoxy composites based on an observational dataset.

A study is also used to predict the Newmark displacement using a gradient-boosted regression tree (GBRT). When compared to other ML approaches, this methodology integrates a succession of regression trees (RTs) into a powerful prediction model and has been proved to be a valuable tool for various data mining

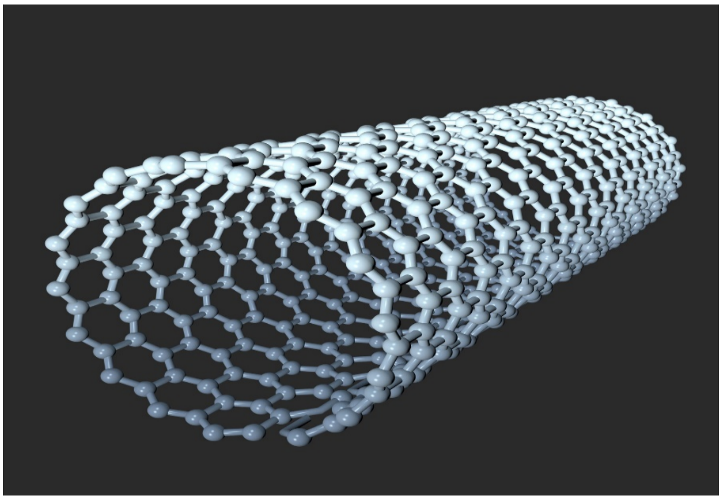

Fig. 1. Carbon nanotubes

challenges. Unlike normal GBRT, the regularization idea was introduced into XGBoost to penalize tree complexity in exchange for enhanced model performance [9].

The aim of this study is to minimize the time it takes to calculate atomic coordinates from hours to seconds. Existing mathematical approaches are acknowledged to be incapable of reducing computation time toward this scale. Hence, Regression tree ensembles are used in this study to achieve this aim. The other sections of this paper as follows; section two gives an introduction to Carbone nanotubes, materials and methods of this study are stated in section three. Finally, results and discussion are explained in section four and the conclusion is stated in section five.

Carbon nanotubes can be considered of as a cylinder-shaped sheet of graphene and these enigmatic formations have generated considerable intrigue in recent years, and considerable research has been devoted to their interpretation. Physical qualities are continuously being discovered and debated at the moment. It is possible to make carbon nanotubes as thin as 1 or 2 nanometers, which is a real example of nanotechnology. They are chemically and physically manipulable molecules that can be of great benefit. Materials research, electronics, chemical processing, energy management, and a slew of other industries benefit from their use [10]. The Fig. 1 illustrates the carbon nanotubes [11].

Carbon nanotubes with a variety of structures have been identified ever since. They are primarily classified as single-walled (SWNTs) or multi-walled carbon nanotubes (MWNTs) based on the number of graphic shells. Fullerene cages are used to seal the ends of SWNTs, which are defined as long tubes made from a single graphene sheet rolled into a 1-nanometer-diameter cylinder. The curvature of the surface is created by the fullerene structures, which alternatively have five hexagons and one pentagon [11]. The Fig. 2 shows the classification of carbon nanotubes [12].

Carbon nanotubes' sidewalls are comprised of

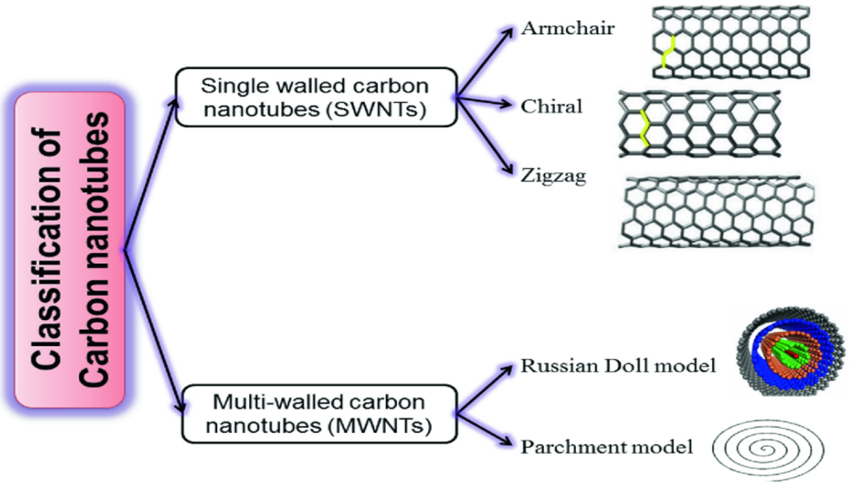
Classification of Carbon nanotubes
Single walled carbon nanotubes (SWNTs)
Armchair
Chiral
Zigzag
Multi-walled carbon nanotubes (MWNTs)
Russian Doll model
Parchment model

Fig. 2. Classification of Carbon nanotubes

Table 1. Samples of dataset

No	Chiral indice n	Chiral indice m	Initial atomic coordinate u	Initial atomic coordinate v	Initial atomic coordinate w	Calculated atomic coordinates u'	Calculated atomic coordinates v'	Calculated atomic coordinates w'
1	2	1	0,67900	0,70131	0,01703	0,721039	0,730232	0,017014
2	2	1	0,71729	0,64212	0,23131	0,738414	0,65675	0,232369
3	2	1	0,48933	0,30375	0,08846	0,477676	0,263221	0,088712
4	2	1	0,41395	0,63299	0,04084	0,408823	0,657897	0,039796
5	2	1	0,33429	0,54340	0,15989	0,303349	0,558807	0,157373

graphene sheets, which are composed of hexagonal cells stacked on top of each other. Pentagons and heptagons are sidewall flaws in other polygon structures. SWNTs with diverse architectures and qualities can be made with different rolling directions on the cylindrical sidewalls. Cylindrical symmetry limits the number of feasible ways for creating seamless cylinders to a tiny number, which are characterized by chiral vectors with constant coefficients (n, m). The electrical properties of a nanotube are directly influenced by its structural architecture. A nanotube is said to be "metallic" (very conductive) when (n – m) is more than 3, otherwise it is a semiconductor. In contrast to other designs, the nanotube in the Armchair is always metallic. [10, 11].

MATERIALS AND METHODS

Datasets and Experimental setup

This study made use of publicly available internet benchmark datasets. The datasets, which has total of (10721) instances with (8) attributes, are provided by the UCI Repository of Machine Learning which has been developed with the aim of assisting the machine-learning (ML) groups in conducting precise research on machine learning techniques [13]. The dataset has no missing values and Table 1 shows a sample of this dataset.

Moreover, this study makes use of MATLAB (R2018b) which is a platform for programming to analyse the data and create models. It contains apps which show the engineers and scientists how learning approaches interact with their data till they get the desired results. The app of regression learner which was obtainable since 2017 is used in this paper to generate the results. It helps users to grasp the correlation between parameters and numerical outcomes. Essentially, this app allows you to design regression models directly and then evaluate their accuracy and performance [14].

Regression Tree Ensembles

Regression tree ensembles are predicting approaches that incorporate many regression tree techniques in a weighted combination merging sundry techniques to improve the performance of the predicting models. The ensemble model's primary idea is that clustering of prosaic classifiers

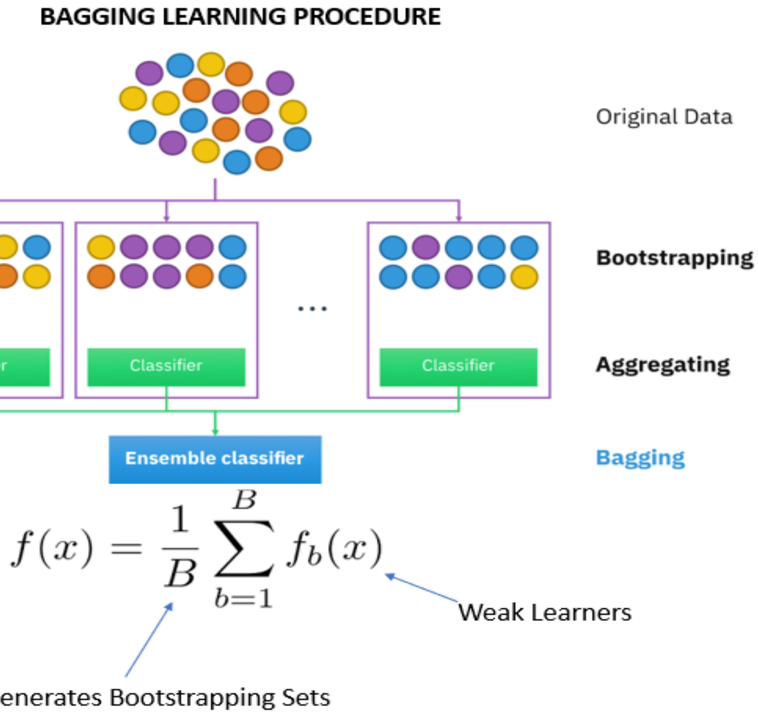

Fig. 3. Bagging ensemble approach

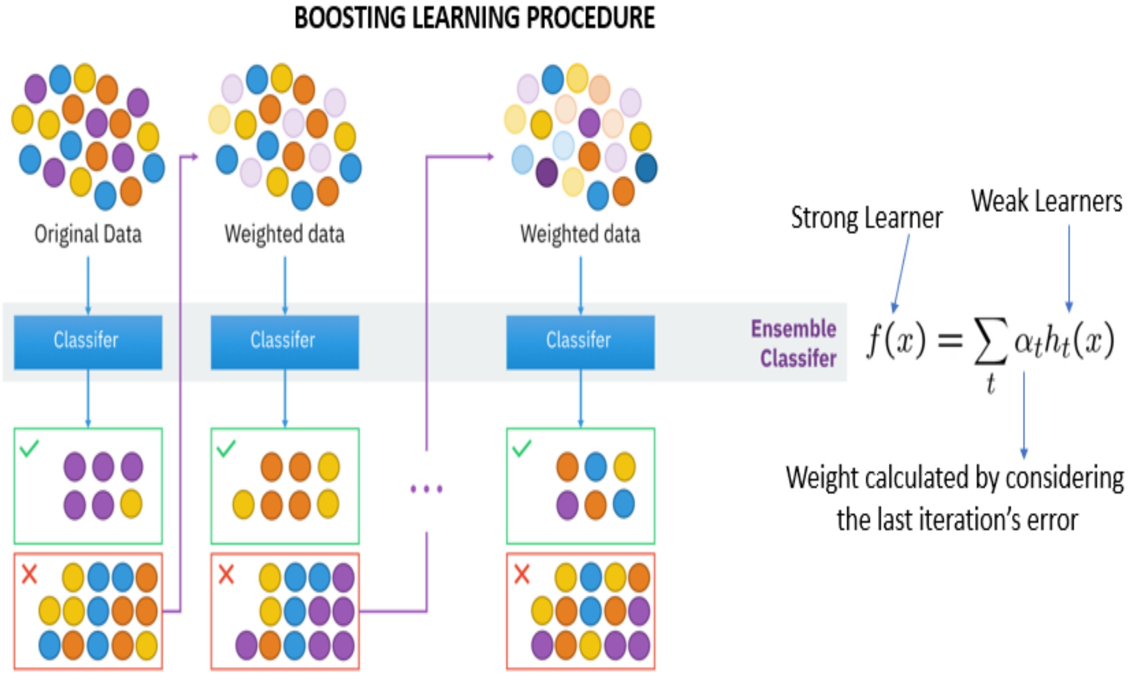
BOOSTING LEARNING PROCEDURE
Original Data
Weighted data
Weighted data
Classifer
Classifer
Classifer
Ensemble Classifer
Strong Learner
Weak Learners
$f(x) = \sum_t \alpha_t h_t(x)$
Weight calculated by considering the last iteration's error

Fig. 4. Boosting ensemble approach

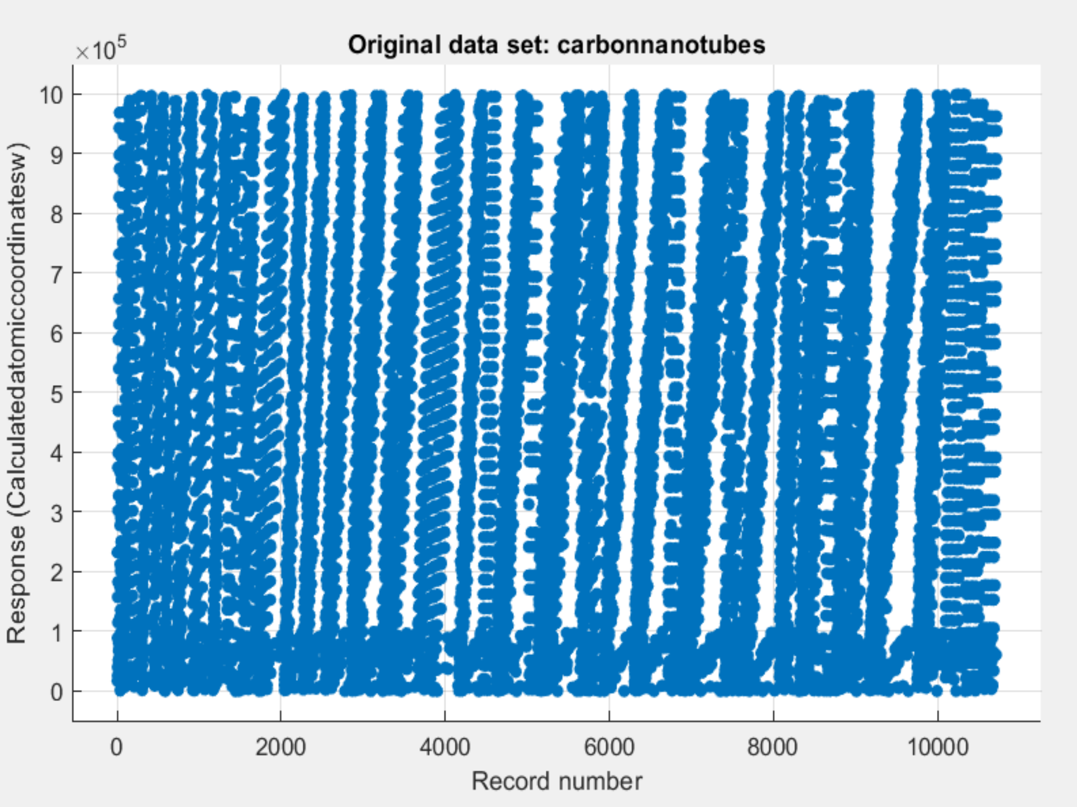

Fig. 5. Carbon nanotube dataset

forms an alliance to create a robust classifier for better prediction. However, bagging which is parallel and boosting which is sequential are two strategies of the ensemble methods. Fig. 3 illustrates Bagging ensemble approach [15].

Furthermore, bagging incorporates all of the learners' outputs and aggregates the prediction of each one by averaged their outputs. In addition, it attempts to eliminate the inconsistency of learning methods by mimicking an approach identified prior to the application of a specific training set. Rather than randomly selecting new training datasets each time, the original dataset is modified by discarding some samples and combining the others.

To generate new datasets, a stochastic sample is placed to instances from the original datasets with replacement. The acquired datasets through oversampling differ from one another, but they are not self-reliant since one dataset serves as the foundation for others. Even though, it appears that bagging produces a combined model which normally outperforms the single model generated from the native training data that is never considerably worse [16, 17]. Otherwise, Fig. 4 shows the Boosting approach [15].

Another ensemble strategy is boosting which makes a group of models so the learners are taught sequentially with initial learners matching simple classifiers to data and afterwards assessing data for faults. Thus, we suit subsequent trees with the purpose of solving for net error from the previous tree at each step. This approach uses aggregating to integrate the outcomes of every classifier. The error of the model is calculated using the weights of samples and the learning model will focus on a certain group of high-weight samples. As a result, boosting approach begins by giving all training data samples the same weight, then pushing the learning model to develop a model for this

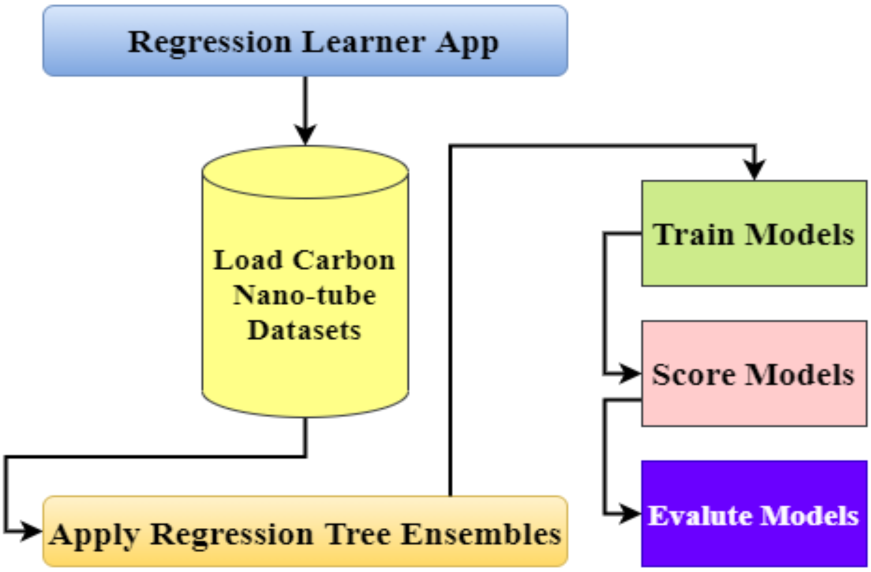
Regression Learner App
Load Carbon Nano-tube Datasets
Train Models
Score Models
Evalute Models
Apply Regression Tree Ensembles

Fig. 6. Flow diagram of the methodology

Table 2. Performance of prediction classifiers

Classifier	RMSE	R-Squared	MSE	MAE	Training Time
Bagged Trees	1.999	0.58	3.998	1.247	7.156 Sec
Boosted Trees	1.952	0.60	3.811	1.348	11.523 Sec

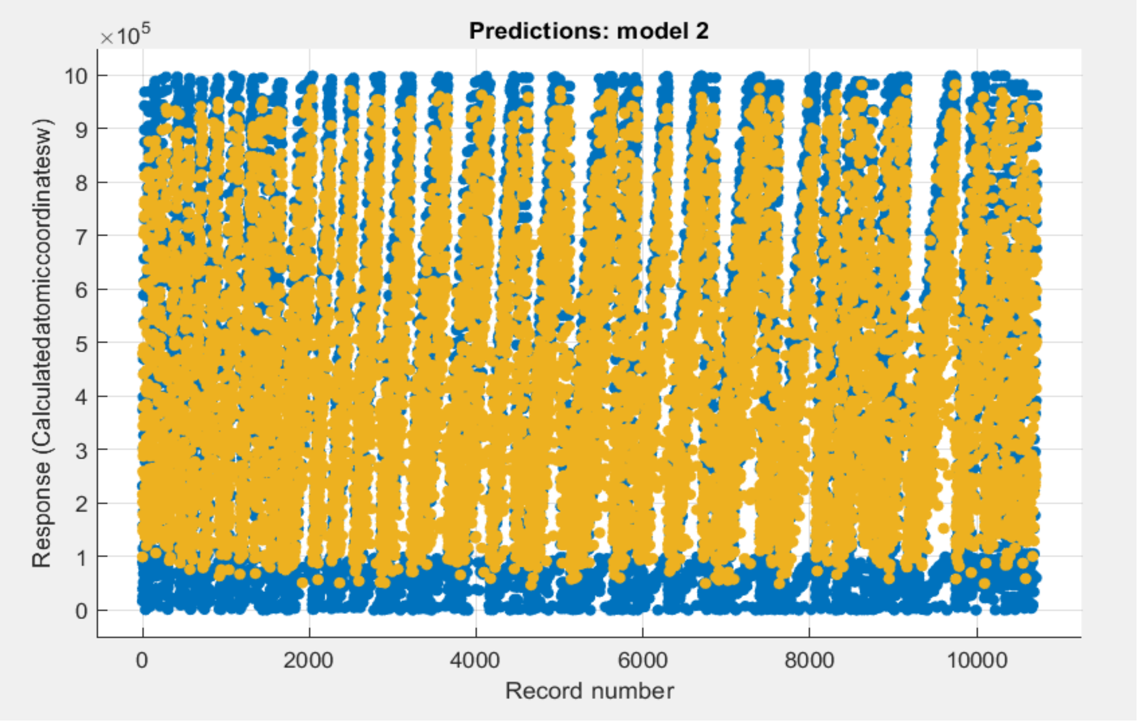

Fig. 7. True and predicted data

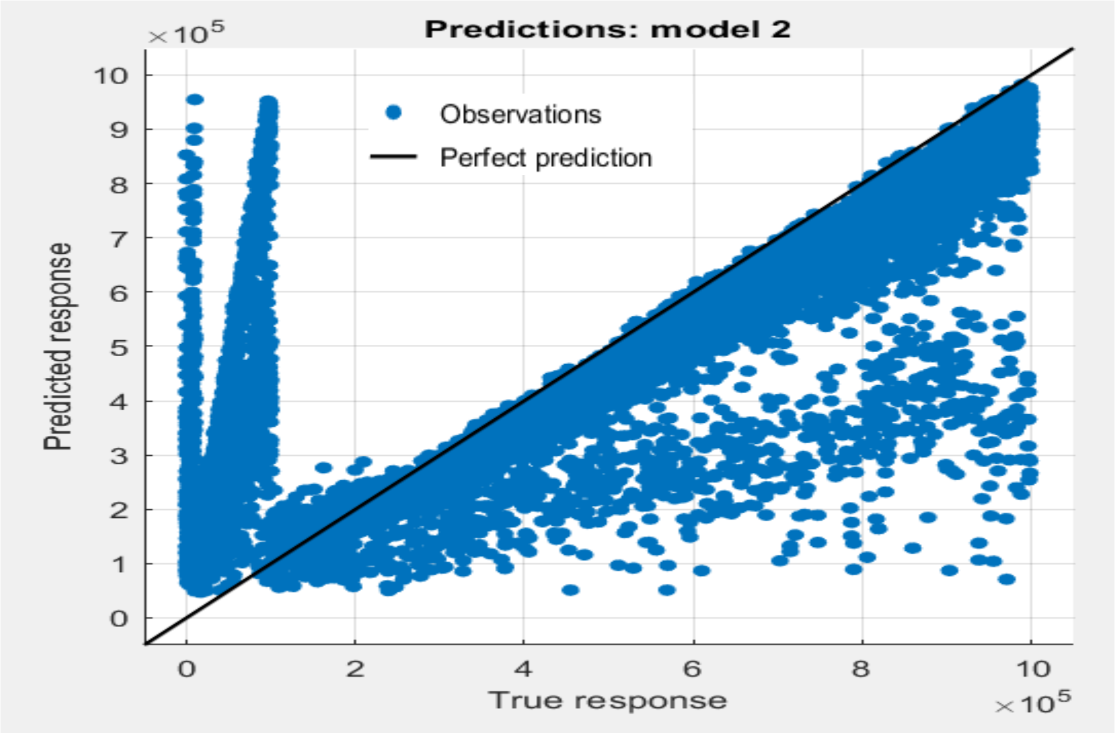

Fig. 8. Actual versus predicted plot

data and re-weighting each sample based on the model's output [16].

Implementation

This study makes use of MATLAB (R2018b) which is a platform for programming to analyses the data and create models. Moreover, we have used two strategies of the ensemble methods which are bagging and boosting. Then, we have applied them to carbon nanotubes dataset. The dataset which has total of (10721) instances with (8) attributes, are provided by the UCI Repository of Machine Learning and has no missing values. Fig. 5 bellow illustrates the original dataset of carbon nanotube when imported into MATLAB.

As mentioned earlier, ensembles models (bagging and boosting) were used in this study for atomic coordinate prediction of carbon nanotubes. In view of the varying portion of training datasets, very few learning methods were used with the aim of finding the best learning plan for the experiment. The Fig. 6 illustrates the methodology of this paper.

The characteristics of these methods were as follow. The minimum leaf size of the bagged approach was (8) and the number of learners was (30) throughout the experiment. On the other hand, the minimum leaf size of the boosted approach was (8), the number of learners was (30), and learning rate was (0.1) throughout the experiment. The K-Folds cross validation which is a technique for resampling the data were set to (5) and principal component analysis (PCA) was set to (off) on both models for better comparison.

RESULTS AND DISCUSSION

This section introduces a comparative analysis of the two classifiers we used in this paper. We used four performance metrics to evaluate these models. Firstly, the Root Mean Square Error (RMSE) which is a commonly used metric for comparing the predicted values with the values observed by a classifier. Secondly, R-squared that is the percentage of variance in the dependent variable that can be predicted by the independent variable. Thirdly, Mean Squared Error (MSE) which is the difference in average squared between both the actual and estimated values. Finally, Mean Absolute Error (MAE) which is the range of error units between actual and predicted values.

Table 2 illustrates the performance of prediction for the two classifiers based on the evaluation

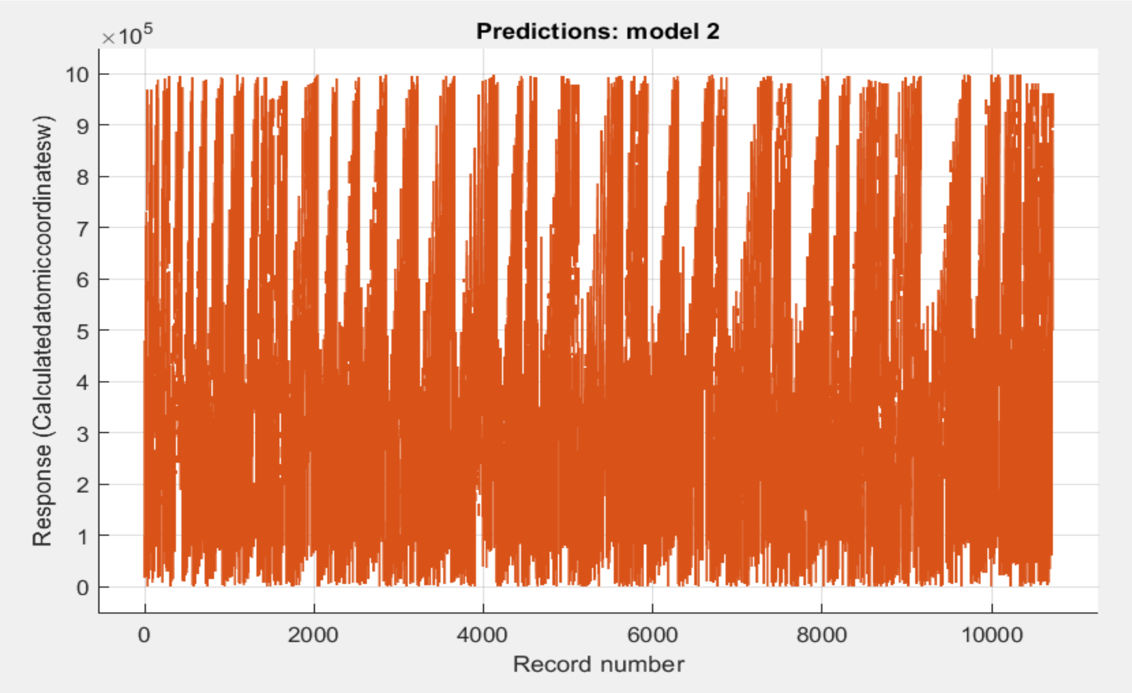

Fig. 9. Error plot of bagged trees

metrics we used in the paper. This performance is based on fivefold cross validated Carbone nanotubes dataset. Th outcomes of classifiers are compared and analyzed.

The Figs. 7-9 show the outcomes of the bagged trees in terms of true (blue) and predicted (gold) data and the actual versus predicted plot and the errors as well.

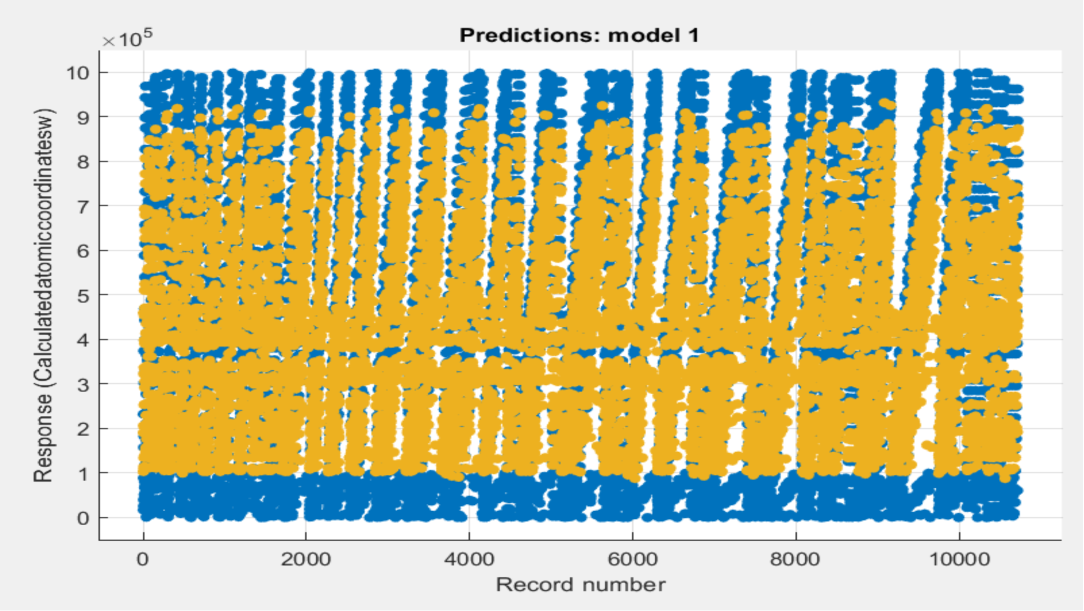

Fig. 10. True and predicted data

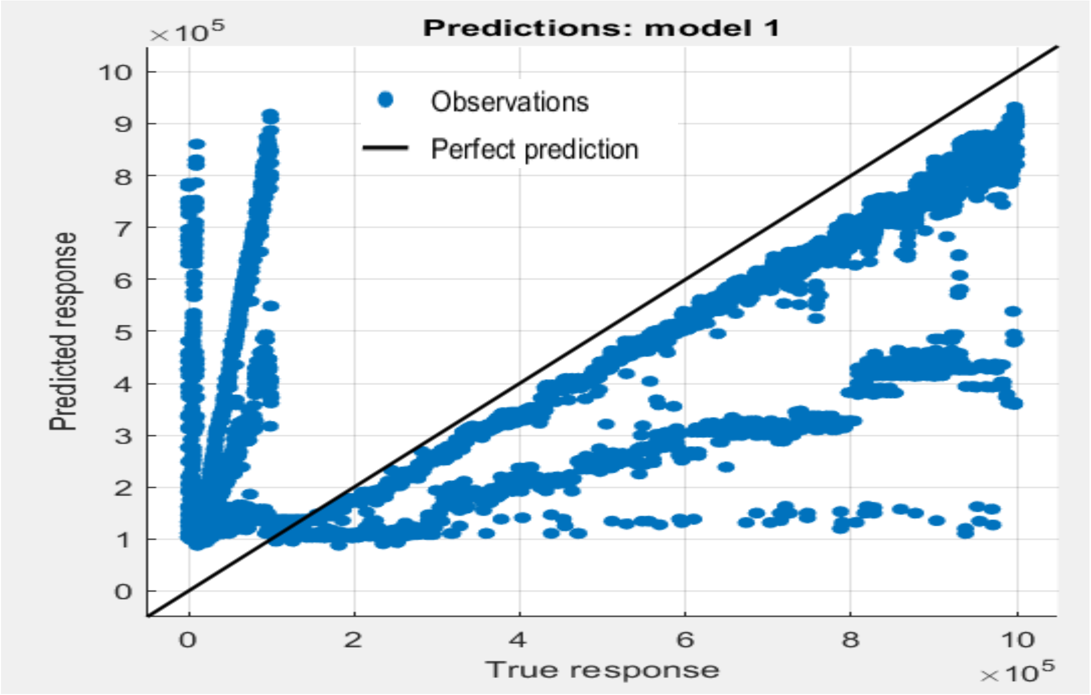

Fig. 11. Actual versus predicted plot

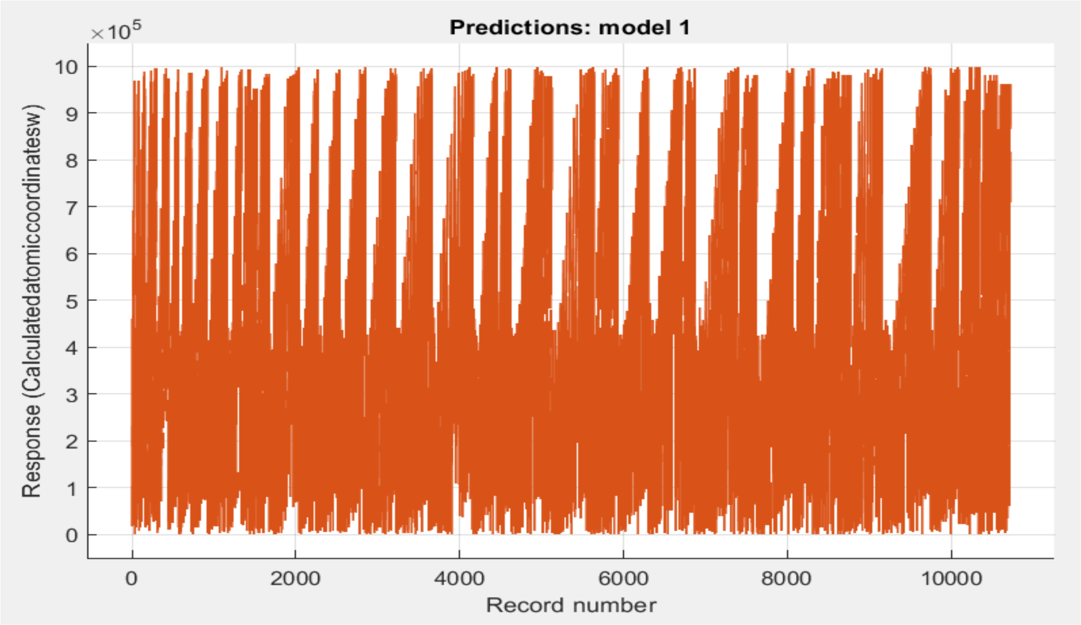

Fig. 12. Error plot of boosted trees

The Figs. 10-12 show the outcomes of the boosted trees in terms of true (blue) and predicted (gold) data and the actual versus predicted plot and the errors as well.

The use of machine learning approaches has proven its effectiveness in prediction of problems in different fields. In particular, bagged and boosted trees have proven their ability in this paper to predict carbon atom coordinates and the results which obtained by them are competitive compared to the study in [2] as an example. Furthermore, it is notable from the Table 2 and Fig. 8 that the bagged trees outperform the boosted trees as the MAE is lower, which means that the range of error units between actual and predicted values are less than boosted trees and this indicates how exact the predictions are in comparison to the desired values. However, R-Squared in boosted trees is higher than bagged trees which means is acceptable as the high value is desirable in this metric. Furthermore, there is a slightly difference between bagged and boosted trees and due to their excellent accuracy and minimal calculation time, the bagged and boosted trees are preferred for computing carbon atom coordinates compared to traditional methods.

CONCLUSION

The use of Regression tree ensembles in predicting atomic coordinate of Carbone nanotubes is proposed in this paper. Carbone nanotubes datasets are used to perform these classifiers. The dataset has a total of (10721) instances with (8) attributes (five as inputs and three as outputs) and it has no missing data. Furthermore, different performance metrics are computed to assess the classifiers we used. The outcomes show that there is a slightly difference between bagged and boosted trees in their performance and at the same time they are preferred for carbon atom coordinates prediction due to their excellent accuracy and minimal calculation time. The actual atomic coordinates can be obtained in minutes or seconds instead of days using these anticipated atomic coordinates as early coordinates for the simulation tool. Future studies can be divided into two categories: The first is to raise the number of features in the dataset. The second category might be to predict atomic coordinates using new machine learning approaches.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interests regarding the publication of this manuscript.

REFERENCES

1. Yamacli S, Avci M. Simple and accurate model for

voltage–dependent resistance of metallic carbon nanotube interconnects: An ab initio study. Phys Lett A. 2009;374(2):297-304.

2. Acı M, Avcı M. Artificial neural network approach for atomic coordinate prediction of carbon nanotubes. Appl Phys A. 2016;122(7).
3. Frhlich R, Rpzany A, Riedmller J, Bolz A, Schaldach M. Electroactive coating of stimulating electrodes. J Mater Sci Mater Med. 1996;7(7):393-397.
4. Quantum-Mechanical Ab-initio Calculation of the Properties of Crystalline Materials. Lecture Notes in Chemistry: Springer Berlin Heidelberg; 1996.
5. Molecular Dynamics Simulation of Ceramic Matrix Composites Using BIOVIA Materials Studio, LAMMPS, and GROMACS. Molecular Dynamics Simulation of Nanocomposites Using BIOVIA Materials Studio, Lammps and Gromacs: Elsevier; 2019. p. 227-258.
6. Aci M, İnan C, Avci M. A hybrid classification method of k nearest neighbor, Bayesian methods and genetic algorithm. Expert Systems with Applications. 2010;37(7):5061-5067.
7. Cheng G, Wu H, Qiang X, Ji Q, Zhao Q. Graphene Field-effect Transistor Modeling Based on Artificial Neural Network. Proceedings of the 2015 International Conference on Mechatronics, Electronic, Industrial and Control Engineering: Atlantis Press; 2015.
8. Cheng WD, Cai CZ, Luo Y, Li YH, Zhao CJ. Mechanical properties prediction for carbon nanotubes/epoxy composites by using support vector regression. Mod Phys Lett B. 2015;29(05):1550016.
9. Chen T, Guestrin C. XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016/08/13: ACM; 2016. p. 785-794.
10. Pournaghshband V, Pournaghshband H. Appending Security Theories to Projects in Upper-Division CS Courses. Computer Science and Information Technology Trends; 2021/12/18: Academy and Industry Research Collaboration Center (AIRCC); 2021. p. 27-37.
11. Carbon nanotube composites. Carbon Nanotube Science: Cambridge University Press; 2009. p. 227-246.
12. Hasnain MS, Nayak AK. Classification of Carbon Nanotubes. SpringerBriefs in Applied Sciences and Technology: Springer Singapore; 2019. p. 11-15.
13. Table 3: Health datasets from UCI machine learning repository. PeerJ.
14. Source code 1. eLife Sciences Publications, Ltd.
15. Ensemble Methods: Bagging and Boosting. Machine Learning: Cambridge University Press; 2022. p. 163-188.
16. Witten IH, Frank E, Hall MA. Embedded Machine Learning. Data Mining: Practical Machine Learning Tools and Techniques: Elsevier; 2011. p. 531-538.
17. Jukic S, Saracevic M, Subasi A, Kevric J. Comparison of Ensemble Machine Learning Methods for Automated Classification of Focal and Non-Focal Epileptic EEG Signals. Mathematics. 2020;8(9):1481.