Exploring Machine Learning Algorithms and Their Applications in Materials Science

[1]Chinua Obasi and [2]Oluyemi Oranu
[1,2]University of Nigeria, Nsukka 410105, Enugu, Nigeria.
[1]obasic522@hotmail.com, [2]oluyemi.oranu@unn.edu.ng

Correspondence should be addressed to Chinua Obasi : obasic522@hotmail.com

Article Info
Journal of Computational Intelligence in Materials Science (https://anapub.co.ke/journals/jcims/jcims.html)
Doi: https://doi.org/10.53759/832X/JCIMS202402003.
Received 28 November 2023; Revised from 20 January 2024; Accepted 31 January 2024.
Available online 16 February 2024.

Abstract – The traditional methods employed in the investigation of new materials, specifically the empirical and density functional theory (DFT) approaches, are insufficient to satisfy the requirements of modern materials science. This can be attributed to the prolonged development cycles, suboptimal efficiency, and exorbitant costs. The utilization of machine learning (ML) is a common practice in material detection, analysis, and design owing to its exceptional predictive capabilities, efficient data processing, and swift development cycle. This can be attributed to its relatively low computational expense. This paper provides an analysis of the essential operational procedures that are involved in the analysis of material properties using ML techniques. Furthermore, the present study provides a summary of the recent utilization of ML algorithms in diverse established domains of materials science, along with a discussion on the requisite improvements for their widespread implementation. The utilization of ML has been widely implemented in various fields of materials science. This paper offers an academic analysis of the paradigms of ML in the context of materials science. The article provides a clear and comprehensive overview of the essential steps involved in data processing, which encompass sample construction, data modelling, and model evaluation. The present manuscript presents a comprehensive survey of the application of ML methodologies in the domain of material science.

Keywords – Materials Science, Machine Learning, Artificial Intelligence, Material Analysis, ML Algorithms.

I. INTRODUCTION

The scientific pursuit of artificial intelligence gave rise to Machine learning (ML). In the 1950s, various symbolic approaches were utilized to address the matter of knowledge acquisition by machines. Following this, a comprehensive investigation was carried out on connectionist methodologies, encompassing neural networks and perceptrons. Subsequently, several methodologies based on statistical learning theory (SLT) were presented, such as decision trees (DTs) and support vector machines (SVMs). Currently, there is an increasing fascination with diverse innovative machine methodologies, such as deep learning, for the objective of scrutinizing vast datasets, which has attracted the consideration of both academic and industrial domains. The term "machine learning" pertains to a methodology that facilitates the automation of the analytical model construction process. Machine learning (ML) is a computational methodology that utilizes iterative algorithms to extract underlying patterns from data, allowing computers to uncover hidden insights without the need for explicit programming of search parameters.

The materials science field employs both computational and experimental techniques to investigate the materials' characteristics and composition. Subject matter experts (SMEs) with significant experience spanning multiple years or even decades utilize their knowledge to create innovative materials that demonstrate the desired structure-property relationships. Over the past decade, the domain of ML has undergone substantial growth and has more recently penetrated scientific disciplines such as materials science, physics, healthcare and astronomy. Presently, scholars in the domain of materials science are employing different approaches of ML in order to extract valuable insights from pre-existing computational and experimental data. The aforementioned methodology is being utilized in order to conduct material characterization, predict molecular properties, expedite simulations, discover new materials, and generate models. The swift advancement of material assessment technology has led to a surplus of data that exceeds the analytical capabilities of domain specialists. The proliferation of literature, including review and perspective articles, concerning the employment of ML in the realm of material science is suggestive of the increasing importance of ML as a mechanism in this area.

The extensive utilization of ML and its ubiquitous integration in field of scientific study has introduced multiple resources, which are easily obtainable for those who wish to embark on their journey in this domain. These resources have

the potential to appear in either of two forms. Commonly encountered resources in the domain of ML are generally broad in nature and take the form of instructional literature or multimedia, web-based curricula focused on ML, and ML certifications. These models often employ openly accessible datasets and employ a generalized ML approach that is well-suited for educational purposes and easily understood by inexperienced students. It is a prevalent issue that sources which employ curated and sanitized public datasets often fail to demonstrate the necessary techniques for accommodating the complexities inherent in domain-specific scientific data. Small and medium-sized enterprises (SMEs) have the option to access alternative sources of ML resources through the dissemination of knowledge by experts in their respective research domains. The aforementioned resources include review articles, research publications, publicly accessible datasets, ML toolkits, and software packages. The resources mentioned above are considered to be more effective than traditional educational materials because they enable a more thorough understanding of the context and provide a perspective that is familiar to the expert in the subject matter.

Numerous scholarly articles on ML in the field of material science, such as such as [1], have focused on particular phases of ML pipeline, including modelling techniques and feature representation. On the other hand, certain articles, such as [2], focus on a specific set of issues within a given field, such as utilizing image-based methods in the realm of material science. The aforementioned articles offer significant perspectives on the comprehensive investigations being carried out in the realm of material science through the utilization of ML. The analyzed research publications presented in these articles illustrate that the results are the product of a multifaceted decision-making procedure that is inherent to the utilization of ML in a new problem area. The process of decision-making is implicit in the existing body of published literature. Numerous scholarly articles have been disseminated regarding the most effective methodologies for integrating ML techniques into the field of material science. Nevertheless, the existing body of literature does not provide explicit documentation of the process of decision-making.

The present manuscript provides an examination of the ML paradigms employed within the domain of materials science. The essential steps involved in data processing, such as constructing a sample, creating a data model, and evaluating the model, are thoroughly defined and analyzed. The present study offers a comprehensive survey of the application of ML methodologies within the domain of material science. The current manuscript is structured in the subsequent manner: Section II focuses on the paradigms of ML in materials science, while Section III discusses the applications of ML in materials sciences. These applications include material discovery and component prediction. Lastly, Section IV presents a conclusion as a well as future research directions.

II. PARADIGMS OF ML IN MATERIALS SCIENCE

One commonly accepted definition of ML is represented by the notation $< P, T, E >$, where P, T, and E correspond to performance, task, and experience, respectively. According to Raymond [3], the primary inference is that programmes can acquire knowledge from experience E concerning a particular task T and based on performance measure denoted by P. The program's capacity to perform tasks within T, as evaluated by P, is enhanced as a result of its experience E. Typically, the development of a ML framework is deemed necessary in the context of utilising ML techniques to tackle a specific issue within the domain of materials science. The overarching model of these ML structures is presented as follows:

$$Algorithm + Sample + Goal = Model \tag{1}$$

The ultimate aim of the inquiry is commonly expressed as an objective function that represents the stated problem. The term "sample" denotes a distinct population subset, which has been chosen for analysis in a predetermined manner. The process of data preprocessing, which involves data cleansing and feature manipulation, is commonly applied to change the initial set of data into sample data. The procedure of data cleansing encompasses identifying and eliminating deficient, erroneous, imprecise, and immaterial elements of the data. The data that are deemed inappropriate or lacking in refinement are subsequently substituted, altered, or expunged in accordance with the appropriate protocols. The process of feature engineering involves a range of methods, such as feature construction, feature construction, feature selection, feature learning, and feature extraction. These techniques rely on domain knowledge to generate meaningful features that facilitate the efficient operation of ML algorithms.

Feature engineering is a vital component in the execution of ML, and it presents significant difficulties and expenses. As stated by Mohan, Neogy, Seth, Garg, and Mittal [4], the Algorithm is a self-contained and comprehensive sequence of operations that includes both ML algorithms and model optimization algorithms. Decision Trees (DT), Artificial Neural Networks (ANN), Support Vector Machines (SVM), are commonly employed ML algorithms. The optimization of models is predominantly accomplished by employing simulated annealing algorithms (SAAs), particle swarm optimization (PSO), and genetic algorithms (GAs) algorithms. The Model refers to a depiction of a particular system that employs computational principles and communicates with the algorithm obtained from the Sample.

ML Steps in Materials Science

The diagram depicted in **Fig 1** portrays the methodology involved in creating a ML system, consisting of three discrete phases: sample generation, model creation, and model evaluation.

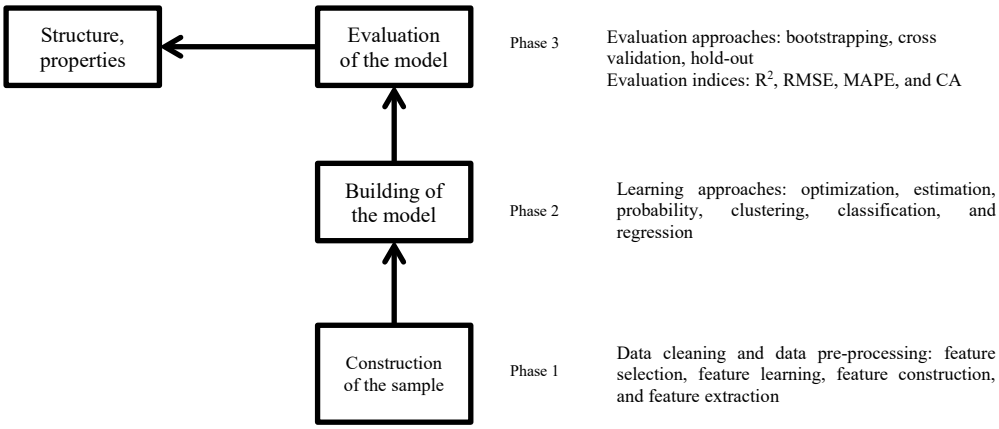

Fig 1. The overarching methodology of ML in the field of materials science.

Sample construction

The process of gathering primary set of data in the materials science field entails obtaining empirical measurements and conducting computational simulations. Data often display incompleteness, noise, and inconsistency, which underscore the importance of data cleaning during sample construction from the initial data. In addition, there are various conditional variables that influence the retrieved sample, and most of these variables are irrelevant to the decision factor/attribute. In the realm of research concerning the forecast of the ionic conductivity of Li, it is noteworthy to acknowledge that even though different external and internal variables may potentially impact the ionic conductivity, about 4 of them are considered to be of utmost significance in comparative experiments. The aforementioned variables encompass mean volume, ion diffusivity, temperature of experimentation, and transition temperature. The aforementioned data has been obtained from a study that has been conducted by Kuanr and Mohapatra [5]. Thus, it is imperative to utilize a suitable feature selection methodology to identify the attribute subset that will be employed in the end simulation.

The utilization of big data has significantly enhanced the information resources available in the field of materials science. Büchi, Festic, and Latzer [6] presented a seven-fold categorization of the effects of big data, which includes the following dimensions: visualization, veracity, variety, value, variability, velocity, and volume. The factors previously mentioned served as obstacles to the employment of data processing approaches in the materials science field. The impact of this aspect on the effectiveness of the ML model is significant and crucial in the field of ML. In general, the data processing procedure comprises two fundamental constituents, specifically feature engineering and data selection.

Data selection

The procedure of selecting data entails a thorough evaluation of diverse factors, including but not limited to the nature, standard, and structure of the data. The utilization of data of superior quality can help to reduce the presence of missing, redundant, or erroneous data. Thus, it is crucial for scholars to obtain information from reliable databases. The Unite States, in 2011, launched an initiative known as Materials Genome to focus on the significance of big data following the advancement of materials science. The initiative fervently promoted the establishment of an enhanced material repository. The domain of computational materials science has made use of various material databases such as the Open Quantum Material Database, Harvard Clean Energy Project, Material Project, AFLOWLIB, Inorganic Crystal Structure Database, and Computational Materials Repository.

Moreover, the application of data mining approach has been employed to retrieve essential scholarly articles pertaining to materials with the aim of improving pre-existing databases. The utilization of ML techniques has been proposed by Punnam, Dutta, Krishnamurthy, and Surasani [7] to train a model using failure data obtained from unsuccessful experiments. This approach is suggested as a means of implementing data processing methodologies in the field of materials science. The scholars utilized empirical data derived from hydrothermal synthesis reactions that were either unsuccessful or suboptimal to develop a ML algorithm that aids in predicting the formation of template vanadium-selenite crystalline materials. The model described above demonstrates enhanced performance in contrast to traditional manual analysis methods. It achieves a predictive accuracy of up to 89% in forecasting the formation aspects on novel inorganic products that employ organic templates.

The field of materials science could be grouped into 4 major categories of data: experimental and simulated material properties encompassing physical, chemical, structural, thermodynamic, and dynamic properties, the data for chemical reaction integrating temperature and rate of reaction, image data including literature-retrieved data, and scanning electronic microscope photographs and images of different materials surfaces. The data that has been presented can be categorized into three distinct classifications: discrete data, which pertains to texts; continuous data, which pertains to vectors and tensors; and weighted graphs. The difficulty in consolidating information from diverse databases is attributed to the

presence of dissimilar data structures and disparate storage sites. Furthermore, the designated data structure is dependent on the specific ML algorithm employed. Therefore, it is crucial to establish a uniform data types and choose a suitable representation of data in order to facilitate ML algorithms in the data processing field. Frequently utilized data representations encompass SMILES notation, fingerprint, Coulomb matrix, and weighted graph.

Feature engineering

After the data selection process, it is imperative to engage in feature engineering, which entails recognizing and extracting pertinent attributes that are applicable to the desired prediction target. The term "feature engineering" pertains to the methodical approach of extracting pertinent and enlightening features from unprocessed data, in order to facilitate the efficient implementation of algorithms. The establishment of the maximum achievable performance threshold for a ML model is frequently dependent on a crucial factor, specifically, its overall effectiveness.

Traditional ML methods, commonly referred to as shallow learning, require the manual identification and selection of features. ML techniques were utilized by ur Rehman, Pei, and Yasin [8] to explore the characteristics and possibilities of Heusler compounds. The present study conducted a range of experiments to determine 22 discrete characteristics, which encompassed, among other things, the element B group number, the general p-valence electron number, and the disparity in A/B radii. The principal aim of this investigation was to augment the capacity of computational systems to detect covert associations. Khan, Ali, Ahmad, Hayat, and Pi [9] were able to identify a specific subset of 14 features from an initial pool of 30 features. The aforementioned characteristics comprised the aggregate ionic charge, tolerance factor, electrons in p orbitals, and the summation of p and s orbital radii. The chosen characteristics were employed to educate a ML algorithm with the objective of forecasting unexplored hybrid organic-inorganic perovskites (HOIPs) for the purpose of photovoltaic usage.

However, the analogue process of feature engineering is not a preferred approach. The difficulties arising from the limitations of human perception and understanding make it challenging to identify the most significant features for accurately forecasting the desired result. In addition, the manual process of feature engineering necessitates more computational and labor expenses. The advent of deep learning in recent times has disregarded the necessity for manual feature engineering that has the potential to establish a trend in the domain of ML in the context of materials science.

Model building

Through the implementation of a suitable structure and a substantial quantity of information, it is feasible to develop a framework that can be employed for the examination of substances. The procedure of modeling entails the meticulous selection of appropriate algorithms, the training of these algorithms using pertinent training data, and the production of accurate predictions. The discipline of ML could be categorized into 4 primary sub-disciplines, namely unsupervised learning, supervised learning, reinforcement learning, and semi-supervised learning. Supervised learning is a type of ML that entails the presence of a teacher, whereby the training data is annotated with corresponding outputs. Unsupervised learning differs from supervised learning in that the training data utilized does not possess labeled outputs.

In the domain of ML, the technique of semisupervised learning entails the utilization of a subset of data that has been labeled in conjunction with a more extensive collection of unlabeled data to facilitate the training process. In general, the amount of unannotated data that is accessible tends to exceed that of annotated data. Reinforcement learning is a technique that relies on environmental reinforcement signals to evaluate the effectiveness of actions generated by the model, as opposed to providing explicit instructions to the model on how to generate correct actions. This methodology facilitates the enhancement of methods for adapting to the surroundings. Algorithms are readily available to implement the four distinct categories of ML techniques as previously outlined. The aforementioned categories can be broadly categorized into two distinct groups, specifically shallow learning and deep learning. Conventional ML models frequently utilize shallow learning techniques such as support vector machines (SVM), artificial neural networks (ANN), and decision trees (DT), for linear classification.

Support vector machine

The Support Vector Machine (SVM) represents a linear classifier frequently employed for binary classification endeavors. The SVM algorithm is proficient in discerning a hyperplane of N-1 dimensions for a given set of data points that are situated in an N-dimensional space. For instance, consider the classification of 2D datasets. The hyperplane exhibits the capacity to effectively subdivide the training dataset into 2 different groups. If the algorithm encounters data that cannot be identified, it will utilize the classification model mentioned earlier to analyze the dataset. **Fig 2A** displays the methodology employed by a lineal SVM. The SVM was first introduced in 1964 and has undergone substantial progress since the 1990s, resulting in the creation of numerous enhanced algorithms. This algorithm has been employed in diverse fields, including but not limited to facial recognition, text classification, biomedical research, and other pattern recognition tasks.

The utilization of the SVM algorithm for the categorization of compounds linked to a target drug has yielded exceptional classification efficacy. Furthermore, SVM have demonstrated efficacy in the identification of the drug that exhibits the highest similarity to the focus drug in the main screening process. Moreover, the SVM is a suitable method for discerning the connections between structure and property. The zeolite synthesis process involved the use of different

synthesis factor, such as the first gel level, temperature, time, and the reaction process, which were employed by Hikichi et al. [10] as inputs. By utilizing this methodology, the researchers achieved accurate prognostications pertaining to the structural elements and thermodynamic characteristics of the resultant commodities.

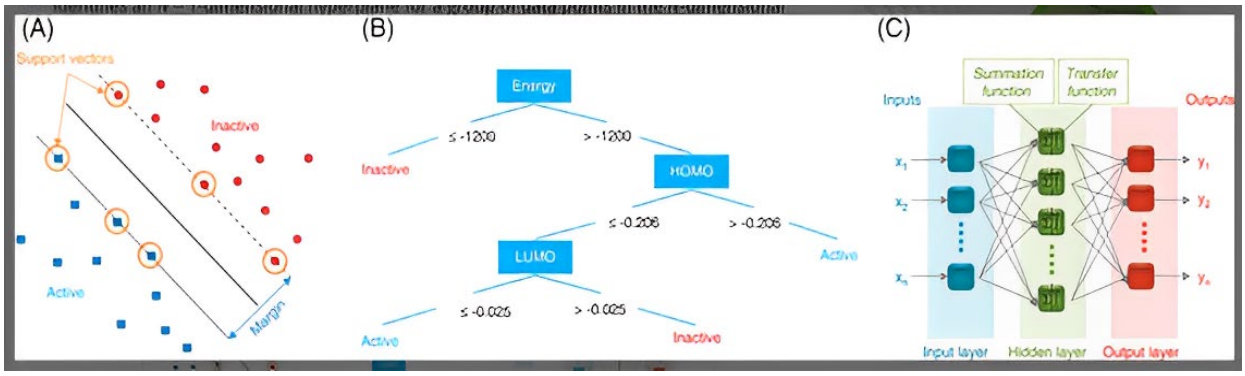

Fig 2. The provided visual representations depict (A) the SVM, (B) the decision tree, and (C) the artificial neural network

Naive Bayes classifier

The topic of the Naive Bayes classifier has been widely researched since the 1950s. The system consists of a collection of straightforward probability classifiers, which relies on the Bayes theorem, based on the assumption of significant independence among the features. The procedure for training classifiers is not dependent on a single algorithm, but rather a set of algorithms that function on the assumption that every characteristic of a particular sample is autonomous of the rest. In the process of object classification, the category with the highest probability is typically assigned to the object, assuming that the probabilities for each category have been acquired. The maximum probability estimation technique is frequently employed to ascertain the parameter of the naive Bayes model in diverse practical contexts. Consequently, it is feasible to implement the naive Bayes model in situations where the Bayes probability or a different model is not utilized. The Naive Bayes classifier is considered advantageous due to its utilization of minimal parameter estimation, specifically the variance and mean of every variable, from a limited set of sample data to produce forecasts. Consequently, it is commonly utilized to forecast the efficacy of a particular methodology, e.g., fabricated formula for an original compound.

Decision tree

The DT (Decision Tree) method is a viable approach for approximating discrete functions. The methodology employed in this study is a conventional classification approach that seeks to achieve accurate sample categorization by deriving a collection of classification protocols from the sample data. The DT layout is presented in **Fig 2B**. The Iterative Dichotomiser 3 (ID3) algorithm, which forms the foundation for Decision Trees (DT), was created by J. Ross Quinlan in the 1960s. The utilization of the ID3 method was subsequently implemented to augment the C4.5 algorithm via diverse approaches, including its pruning mechanism and attribute selection criteria. The decision tree (DT) methodology generally comprises three fundamental phases, namely, feature selection, DT generation, and the elimination of superfluous DTs. The main purpose of feature selection is to preserve only those features, which provide satisfactory performance of classification, while pruning is intended to simplify and generalize the tree. It is imperative to choose a decision tree that displays minimal inconsistency with the training data and adequate generalizability, as there may be several decision trees, which can efficiently categorize the training dataset.

The utilization of Density Functional Theory (DFT) was implemented by Gao et al. [11] in order to produce novel AB2C Heusler compounds. The training dataset was constructed by utilizing the ASM Alloy Phase Diagram Database, and the Pearson's Crystal Data. The scope of data collection was restricted to the subsequent items: The three components demonstrate thermodynamic stability within their respective phases. It is worth mentioning that the aforementioned components do not encompass noble gases, actinide or radioactive compounds, or hydrogen. Furthermore, they demonstrate a stoichiometric ratio of 1:2:1, which is considered ideal. A collection of twenty-two attributes, encompassing element B group number, and the aggregate p valence number electrons, the radius difference between elements A and B, and the corresponding electronegativities of the elements, were chosen to delineate the characteristics of the Heusler compounds. The random forest methodology was employed to combine the results of various predictors, each of which was trained separately, through the use of multiple decision trees. A subset of the complete dataset is utilized to train Decision Trees for sub-prediction.

Artificial neural network

Huang [12] developed the MP model as a mathematical framework. The MP model by Tartaglione, Bragagnolo, Odierna, Fiandrotti, and Grangetto [13] presented a meticulous mathematical depiction of neurons and network structure, showcasing the capacity of an individual neuron to perform logical operations. This event signified the beginning of ANN, which is a data processing model that comprises interconnected processing units, commonly referred to as neurons. It

exhibits nonlinear and adaptive behaviour. The present text outlines an information processing architecture that is nonprogrammed and adaptive in nature. This architecture employs network transformation and dynamic behaviour to achieve parallel and distributed data processing. The key aim of this system is to replicate the cognitive abilities of the human nervous system in information processing. The Artificial Neural Network (ANN) is a multidisciplinary system that integrates the fields of neuroscience, computer science, and artificial intelligence. **Fig 2C** illustrates the architecture of the ANN.

Artificial neural networks (ANNs) are composed of interconnected neurons that are arranged into layers and possess the ability to represent a variety of objects, including concepts, words, abstract patterns, and features with significance. The Artificial Neural Network (ANN) is comprised of three discrete layers, namely the hidden layer, output layer, and input layer. The function of the input layer is to receive data or signal from the external environment. The ultimate stratum of the system generates the refined outcomes. The intermediate layer, which is positioned between the output layer and the input layer, is not perceptible from outside the network. The software performs calculations by utilizing its basic operation. The input dataset of ANN is concatenated into a novel vector and thus changed to matrix before being being processed by the network.

During the transmission of data across the system, the i^{-th} neuron found in the input layers executes a multiplication operation on the input data utilizing the weight W_{ij}. The resultant output is subsequently conveyed to the j^{-th} neuron situated in the succeeding layer. The weight in the artificial neural network (ANN) is responsible for representing the inter-neuron connectivity of the neural network. The ANN can adjust this weight to enhance the general model performance. The aggregation of the weighted inputs originating from the neurons situated in the hidden layer is integrated with the bias term and subsequently subjected to the activation function prior to being conveyed to the subsequent layer. The final output is projected by applying an efficient function of transformation in the output layers. The key merits of Artificial Neural Networks (ANN) are as follows: The system exhibits a capacity for autonomous learning, an element for associative encryption, and proficiency in swiftly identifying optimal solutions.

Around 40 unique models of neural networks have been proposed, including but not limited to perceptrons, backpropagation networks, self-organizing maps, Boltzmann machines, and the Hopfield network. ANN has been widely utilized in various domains of materials science, including quantum computing, materials property evaluation, and nanomaterial synthesis, among others. Even though shallow learning could produce significant results in different materials science field, it is faced with several challenges. It is important to acknowledge that shallow learning algorithms, despite their capacity to reduce computational costs, are incapable of attaining commensurate levels of accuracy with DFT across various tasks. The utilization of shallow learning algorithms requires the involvement of researchers with domain expertise to manually engineer features that are suitable for the input data. The aforementioned phenomenon results to a reduction in the model's accuracy.

Recent progress in the field of deep learning has resulted to novel advancement in the various applications of data-driven approaches in materials science. As mentioned earlier, the methods of shallow learning, which rely on linear classification and analogue feature extraction, are aptly effective for linear classification-based tasks. Nonetheless, the accomplished dimensions of performance are insufficient in the context of nonlinear classification tasks. Empirical evidence suggests that deep models outperform shallow models in addressing nonlinear tasks. The utilization of a nonlinear cascade processing unit is credited for the extraction of inherent features. The methodology described involves identifying lower-level characteristics to obtain a wider, more abstract depiction of feature classifications.

Currently, deep learning exhibits strong effectiveness across various fields including but not limited to image identification, speech interpretation, comprehension of natural language, biomedical research, and others. Various architectures, including recurrent neural networks (RNNs), convolutional neural networks (CNNs), deep coding networks (DCNs), and deep belief networks (DBNs), have shown remarkable effectiveness in the domains of materials detection, quantum chemistry, and materials design and analysis, within the discipline of materials science. The following section will provide an overview of CNN and RNN. The following methods will not be elaborated on in great detail: The text presents an analysis of three notable deep learning methodologies, specifically DBN, deep stacking network, and deep Q network. These techniques possess the ability to not only recognize and categorize data, but also to produce data. Deep stacking networks are integrated with different blocking networks and can be readily trained through a supervised methodology. Conversely, the deep Q network signifies a foundational algorithm, which merges reinforcement learning and deep learning, thereby facilitating comprehensive learning to an action from perception.

The inadequacy of material databases poses a substantial obstacle to the effectiveness of deep learning in tackling various predicaments in the realm of materials science. Furthermore, the protracted duration necessary for training and the constrained interpretability of deep neural networks (DNNs) could potentially lead to suboptimal performance in comparison to shallow learning methodologies when tackling specific issues. Therefore, the choice of the algorithm for modeling ought to be predicated on the particular task being undertaken.

Convolutional neural network

The Convolutional Neural Network (CNN) is a variation of the Artificial Neural Network (ANN) that functions in a unidirectional manner. The Convolutional Neural Network (CNN) is a computational framework that is based on fundamental principles of both complex and simple visual neuroscience cells. The convolutional neural network (CNN) is capable of accepting images as input without the need for feature extraction and data reconstructions, which are

commonly, involved in traditional image recognition algorithms, by integrating the artificial neural network (ANN) with discrete convolution for image analysis. Kunihiko Fukushima, in 1980, presented the neocognitron, a model of neural networkdesigned to enhance visual pattern recognition. This model is considered a precursor to the CNN. Following multiple endeavors by Wu [14] to initiate the learning of multi-layered networks utilizing diverse techniques, the effectiveness of convolutional neural networks was limited by inadequate computational resources as the network's depth rose. From 2006, the improved processing unit for graphics has become a widely used computational tool, facilitating the progress of Convolutional Neural Networks.

Recurrent neural network

The convolutional neural network (CNN) demonstrates a deficiency in inter-neuronal connections within the same layer, leading to a one-way transmission of information to the hidden layer from the input layer, ultimately resulting in the output layer. As a result, CNNs face a challenge when it comes to processing correlated data. Consequently, it is advisable to utilize a Recurrent Neural Network (RNN) to effectively manage sequential dataset. RNNs are a form of neural networks that maintains a state representing the output of the initial time phase. This state is then combined with the input layer data to determine the present output. The representation of the present condition of a Recurrent Neural Network (RNN) is denoted by S_i, wherein i signifies the count of iterations. The network's input and output are represented by x_i and y_i, correspondingly. The utilization of the aforementioned variables can be employed to compute the present condition of the network.

$$S_i = f(W * S_{i-1} + U * x_i) \tag{2}$$

While f represents a nonlinear element W and U are the parameters that make up the RNN. Based on the aforementioned formula, all RNN layers may share the same set of input parameters. In the context of the hidden layer's memory architecture, the state of the network Si is typically visualized as the memory unit. Each time around, a nonlinear function's input is updated based on the preceding step's output. So, it follows that RNNs need less parameter learning than CNNs do. Machine translation and voice recognition are only two examples of how the Recurrent Neural Network (RNN) has been put to use in the field of Natural Language Processing (NLP). In materials research, it has been suggested that a Recurrent Neural Network (RNN) be used to simulate a similar reward system in order to create new materials with desirable characteristics.

Nonetheless, model bias and model variances are examples of mistakes that could emerge during training. Generally speaking, model bias may be traced back to incorrect assumptions made by the algorithm, whereas model variance can be traced back to the model's sensitivity to subtle changes in the training data. Errors may also occur owing to computational limitations or a lack of data, in addition to the aforementioned causes. It's also important to remember that overfitting might occur sometimes. This issue develops when the hypothesis becomes extremely strict in an effort to preserve consistency, which may lead to poor results for the model when predicting data that differs from the training set. Inadequate modelling sample selection, such as a lack of samples, the use of improper sample selection procedures, or mislabeled samples, may lead to overfitting since it fails to accurately reflect the established classification criteria. Overfitting may also happen if there is a lot of interference in the model, leading the computer to mistake some noise for a feature.

An inappropriate model hypothesis or a failure to meet the circumstances under which the hypothesis is valid may also lead to overfitting. Finally, overfitting may also be caused by a large number of parameters or a complicated model. Unrestrained expansion may lead to growth that is restricted to either little event data or no event data at all in the context of the DT model. Therefore, although such expansion may show a perfect match to the training data, it may not be flexible enough to include additional datasets. In cases when the ANN's decision surface is not unique with respect to the sample data, the approach of back propagation could converge on a critically complex surface of decision. Overfitting may also cause the model to become sensitive to noise or irrelevant details. Thus, model validation is fundamental for minimizing errors and eliminating overfitting.

Model evaluation

A data-driven framework needs to perform admirably not just on the data it was trained on, but also on data it has never seen before. In most cases, the generalization errors of models may be evaluated computationally, and the results used to choose the best model. Testing data must be obtained before a model's discriminative skills may be assessed on a novel set of data. Therefore, the generalization error may be estimated using the testing error received from the test data. Given a single dataset D of m samples, it is feasible to split D into two subsets, S and T, for use in training and testing. Several types of testing and gauging methods may help with this.

$$D = \{(x_1, y_1), \ldots\ldots\ldots (x_m, y_m)\} \tag{3}$$

The dataset D is divided into a test set T and a sample set S in the grasp-out approach, such that:

$$D = S \cap T \; and \; S \cup T = \emptyset \tag{4}$$

Stratified sampling is a sampling technique that preserves categories and is frequently employed to uphold the uniformity of data distribution while minimizing the introduction of further discrepancies. Due to the absence of an optimal solution for determining the relative ratios of S and T, it is common practice to allocate approximately 2/3 to 4/5 of the samples in D to S, while the remaining samples are assigned to T. The approach of cross-validation involves the partitioning of the initial dataset D into k distinct and non-overlapping equal-sized subsets. Every subset Di is obtained through a process of "stratified sampling".

$$D = D_1 \cup D_2 \cup \ldots\ldots\ldots \cup D_k, D_i \cap D_j = (i \neq j)\emptyset \tag{5}$$

Subsequently, the amalgamated set comprising $k-1$ subsets is designed as a training set S, whereas the remainder is allocated for employment as the test set T. The method referred to as leave-one-out cross-validation (LOOCV) is unimpacted by the random set separation when k is the same as the sample number m in D. It should be noted that the aforementioned procedure is iterated for every subset of the data, resulting in a cumulative count of k experiments. Hence, the cross-validation technique could be time-consuming and may not be appropriate for datasets of significant size.

LOOCV offers two key advantages over K-fold cross-validation. First, the model is trained using the vast majority of data in each iteration, producing results that are more consistent with the distribution of the original samples. Lack of chance throughout the experiment ensures that the method and results can be replicated. LOOCV shares the disadvantage of high computational cost with K-fold cross-validation. Implementing LOOCV while working with a large number of raw data samples might be difficult until mathematical evaluation are parallel to minimize the overall computation time.

Models' precision and consistency may be evaluated with the use of validation strategies like the Bootstrap and the Repeating Learning Test (RLT) cross-validation. When validating a model, Randomized Lasso-Type (RLT) cross-validation partitions the dataset in a different way than Leave-One-Out Cross-Validation (LOOCV). As a result, there is a dramatic reduction in the amount of computational complexity. The selection of the test set often relies on the real issue at hand, and determining the ideal quantity of data for model validation may be a difficult process. The generalization error of a model may be estimated with the use of the Bootstrap cross-validation methodology, which employs random sampling to generate estimates. Although the increased computing cost is a drawback, the success of this method minimizing the K-fold cross-validation variance is significant.

According to the bootstrapping sampling approach, the bootstrapping approach integrates constantly copying a sample from D into a novel D' in a random manner until D' has enough m samples. D' represents the training set of data, while DD′ is the test dataset. Since the bootstrapping method keeps the same number of training samples as the original dataset, it is an appropriate strategy when the dataset size is restricted and splitting the training/testing data is difficult. It is possible that estimate bias will be introduced if the bootstrapping method is used since it alters the distribution of the original dataset.

The efficiency of an algorithm is measured by analyzing how well it performs in various scenarios. The model's predictive performance is measured by contrasting the experimental data with the projected results. The criteria for assessment are conditional on the specifics of the topic under consideration. Classification accuracy (CA) is a useful metric for evaluating classification models:

$$\frac{S}{N} = CA \tag{6}$$

where S represents the samples number, which were successfully categorized and N represents the overall sample number. Root mean square error (RMSE), correlation coefficient (R2), and Mean absolute percent error (MAPE), are all utilized to evaluate the models employed to mitigate regression issues, as indicated in expressions (3)-(5).

$$\frac{1}{n}\sum_{i=j}^{n}\frac{y'_i - y_i}{y_i} = MAPE \tag{7}$$

$$\sqrt{\frac{1}{n}\sum_{i=j}^{n}\left(y'_i - y_i\right)^2} = \quad RMSE \tag{8}$$

$$\frac{[\sum_{i-1}^{n}(y1-y)(y'_i-y')^2]}{\sum_{i-1}^{n}(y1-y')^2.\sum_{i-1}^{n}(y1-y)^2} = R^2 \tag{9}$$

where y'_i and y_i signify a corresponding forecasted value and the original value, correspondingly, and y' and y represent the mean of the forecasted value and the original value, correspondingly.

In addition, there exist various alternative metrics that can be utilized to evaluate the efficacy of classification models. These include but are not limited to receiver operating characteristic (ROC) curve, precision, recall, hinge loss, Cohen's kappa, confusion matrix, Jaccard similarity coefficient, logistic regression loss, Hamming distance, coverage error, ranking loss, and label ranking average precision. The utilization of indices of detailed coefficient and variance of determination is common in the context of regression analysis. The mutual information (MI), silhouette coefficient, and Rand index, are frequently utilized metrics for evaluating the efficacy of models in clustering tasks.

III. APPLICATION OF ML IN MATERIALS SCIENCE

Material discovery

Around thirty years ago, John Maddox, the then-editor of the scientific journal Nature, articulated that a persistent challenge in the field of physical science is the inability to predict the structural configuration of fundamental crystalline solids solely based on their chemical composition. Despite the modern progressions, the act of forecasting the crystal model solely according to its composition remains a vital and perhaps the most significant pursuit in the domain of materials science. The rationale behind this assertion is that the development of any rational materials design necessitates a comprehensive comprehension of the crystal model.

However, the process of predicting molecular or crystal models is highly challenging, as the potential configurations of atoms in 3D spaces are vast and complex, resulting in a highly intricate energy landscape. Recent advancements in energy assessment approaches have diversified the "classical" scope of crystal model forecasting approaches to encompass a wide spectrum of solid forms and molecules. Additionally, more intricate algorithms for structure generation and selection, such as simulated annealing, metadynamics, random sampling, evolutionary algorithms, and minima hopping, have been devised in recent times. Nevertheless, these methodologies continue to be computationally intensive as they necessitate a significant amount of force and energy evaluations. The pursuit of novel and enhanced high-performance materials requires a thorough investigation of the composition and structure domain. The potential of ML computations in addressing the challenge at hand is highly promising, owing to the integration of vast quantities of data.

There exist diverse methodologies that can be utilized by ML algorithms to tackle this concern. A plausible initial tactic entails expediting the energy assessment procedure by replacing a first-principle approach with ML algorithms that demonstrate considerably higher velocity, as explicated in the "ML Force Fields" segment. The prevalent approach employed in the domain of inorganic physics in solid-state is commonly denoted as an element prediction. Rather than performing an exhaustive analysis of the structural space for a singular composition, scholars choose to designate a prototype framework and then investigate the composition space to ascertain stable materials. Thermodynamic stability is the fundamental concept in the context being discussed.

The phrase "non-decomposable compounds" pertains to chemical compounds that exhibit no tendency to decompose into separate phases or compounds, even over an indeterminate duration. Metastable compounds, such as diamond, can be synthesized and have become more accessible as a result of advancements in the field of chemistry. Typically, it is more practical to generate and control compounds that exhibit thermodynamic stability. The conventional approach to assessing thermodynamic stability involves evaluating the energetic detachment from the convex hull. In particular instances, ML approaches will evaluate the chances of a particular elements' prevalence in a particular phase without intermediary steps.

Component prediction

The predictive capability of formation energy in determining the stability of a recently synthesized compound is evidently constrained. Preferably, it is ideal to employ the detachment to the thermal stability convex hull. Contrary to energy development, the convex hull detachment considers the alterations of free energy, which happens across potential decomposition pathways. The statement mentioned above is erroneous since our understanding of the convex hull is inherently inadequate. Fortunately, the progress in comprehending the convex hull by discovering more resilient materials has led to a reduced importance of this matter over time. It is a widely accepted convention to conduct initial energy computations based on fundamental principles while assuming zero temperature and pressure conditions. This approach disregards the impact of kinetic factors on stability.

The methodology of Kernel Ridge Regression (KRR) was employed by Rezaei, Amirshahi, and Mahbadi [15] to ascertain the formation energies of an extensive assemblage of elpasolites, comprising a total of two million crystals. The crystals were comprised of elements from the main group, with bismuth being among them, and exhibited a stoichiometry of ABC2D6. It was reported that a sample set with 104 composition showing errors of 0.1 eV/atom. Additionally, a total of 90 novel stoichiometries were forecasted to be found in convex hull. The dataset produced by Telschow et al. [16] consists of DFT computations of 250,000 stoichiometry ABC3 cubic perovskites, integrating all components to bismuth, whereas eliminating lanthanides, and rare gases. After conducting experiments with different ML techniques, it was found that the amalgamation of adaptive boosting and random trees exhibited the most noteworthy level of achievement, resulting in a 0.12 eV/atom mean error. **Fig 3** illustrates that the chemical composition has a significant impact on the prediction error. Furthermore, a methodology that focuses on active learning through pure exploitation was suggested, as explicated in the section entitled "Adaptive Design Process and Active Learning".

Dos Santos, Pich, Back, Smiderle, Dumas, and Moura [17] aim to explore two ternary prototypes composition spaces, i.e. tP10-FeMo2B2, and tI10-CeAl2Ga2, which possess a stoichiometry of AB2C2, in order to identify stable compounds. The research methodology utilized in this study is founded on the approach outlined in the specified reference. The convex hull yielded the discovery of 1893 novel compounds, leading to a computation time reduction of roughly 75%. The tP10 and tI10 compounds were found to have false negative rates of 0% and 9%, respectively, according to the report.

The investigation carried out by Baumgartner, Kropf, Haider, Veeranki, Hayn, and Schreier [18] employed conventional Random Forests (RFs) for the purpose of predicting formation energies. This was achieved by utilizing features derived from atomic properties and Voronoi tessellations. The performance of descriptors was found to be superior to Coulomb matrices and partial radial distribution functions, based on a training set comprising of approximately

30,000 materials. Structural data integration retrieved from Voronoi tessellation did not result in any noticeable improvement in the results obtained for the cohort of 30,000 unique materials used for training. The aforementioned observation is based on the underlying assumption that the dataset in question comprises a finite set of materials that possess the same composition, but display discernible structural attributes. The assertion was substantiated by modifying the training dataset to encompass a notable 400,000 material sourced from open quantum materials set. The findings suggest that the model which integrated structural information demonstrated superior performance compared to the composition-only model, exhibiting a 37% reduction in error.

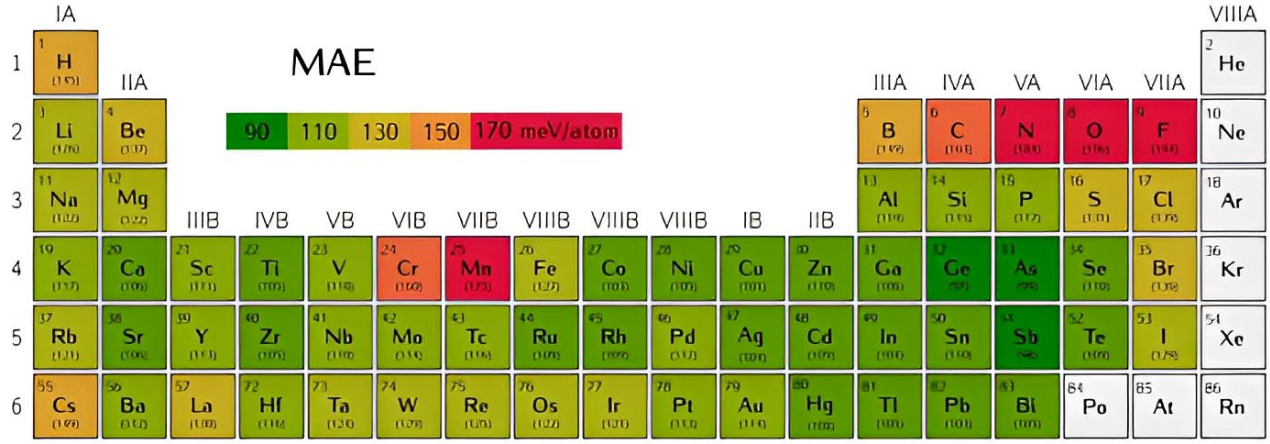

Fig 3. Mean average error, expressed in meV/atom, resulting from the implementation of adaptive boosting in conjunction with extremely random trees

The metric was computed for all perovskites that contain the element under investigation and subsequently averaged. The numerical values enclosed in parentheses represent the precise average error for each individual element. The study conducted by Xia, Zhang, Yuan, Liu, and Ma [19] employed a uniform methodology to investigate quaternary Heusler compounds, leading to the discovery of 53 previously unknown stable frameworks. The model was subjected to training using diverse datasets, which encompassed the comprehensive open quantum material database and exclusively the quaternary Heusler compounds. The research findings indicate that the integration of supplementary prototypes in the training dataset resulted in an improvement of the prognostic efficacy of Heusler compounds. It is fundamental to consider that the utilization of kernel-based methodologies such as Support Vector Machines (SVMs), and Kernel Ridge Regression (KRR) may not be feasible for carrying out research studies that entail voluminous datasets, as a result of their critical computational scaling. The study conducted by Korhonen, Tuppurainen, Asikainen, Laatikainen, and Peräkylä [20] involved the utilization of classification methodologies and diverse regression to scrutinize a dataset consisting of 2150 A1, xA′xB1, yB′yO3 perovskites.

The aforementioned materials exhibit potential utility as cathodic components within solid oxide fuel cells operating at elevated temperatures. A total of 238 elemental characteristics were employed as attributes in the various approaches used. The study's findings suggest that trees that were randomly assigned in an exceptional manner demonstrated the highest efficacy as classifiers, with 0.93 accuracy and 0.88 F1 score. Regarding regression, it was observed that both Kernel Ridge Regression (KRR) and Extremely Randomized Trees (ERT) exhibited remarkable performance, as their respective mean average errors were found to be less than 17 meV/atom. The constrained nature of the elemental composition space poses a difficulty in the comparison of the errors presented in this study with those of other works.

The topic of stability in oxide-perovskite systems is addressed by Dey, Das Sharma, and Mukhopadhyay [21]. The formation energy of non-mixed perovskites was predicted by Van Gompel et al. [22] with a 30 meV/atom mean average error. The achievement was attained by utilizing neural networks that solely depended on the fundamental concepts of electronegativity and ionic radii. However, the dataset at hand was limited to a modest quantity of 240 compounds, which were utilized for the purposes of training, cross-validation, and testing. Shen et al.' [23] study demonstrated that mixed perovskites, which possess two distinct components of B-site and A-site, display comparable error rates. The investigation yielded 9 meV/atom and 26 meV/atom mean average errors for garnets that were not mixed and those that were mixed, respectively, and had a composition of C3A2D3O12. The reduction of the previously mentioned error to 12 meV/atom was achieved by Beglari, Goudarzi, Shahsavani, Arab Chamjangali, and Mozafari [24] through the utilization of structural descriptors and the imposition of constraints on the mixing process exclusively within the C-site. After conducting a comparative analysis with previous research works, it is apparent that the errors observed in this study are considerably minimal. The disparity in the total number of compounds examined between [25] and [26] is noteworthy, given that the former focuses on a mere 600 compounds, while the latter delves into roughly 250,000 compounds. This discrepancy provides a straightforward and unambiguous justification. To clarify, the subject matter exhibits a significant degree of intricacy, exceeding a scale of 100.

Convolutional Neural Networks (CNNs) have been developed by Cheng, Wang, Yang, and Nakano [27] with the aim of identifying crystallographic phases. The convolutional neural networks (CNNs) utilized in this specific context are identified as crystal graph convolutional neural networks (CGCNNs). The concept of Materials Graph Networks was initially presented by Gurnani, Kuenneth, Toland, and Ramprasad [28], while the notion of Message Passing Neural Networks was put forth by Li and Cheng [29]. The section titled "Basic Principles of ML - Features" elaborates on the ability of ML models to predict formation energies. Consequently, they can be employed to expedite component prediction. Thus far, all methods for predicting constituents in this domain have relied on training data derived from first-principle calculations. The scarcity of data regarding finite temperature conditions can be ascribed to the significant computational expenses associated with it. The available data primarily pertains to conditions of zero temperature and pressure, thereby disregarding the influence of kinetic factors on stability. In addition, it is important to mention that metastable compounds, such as diamond, demonstrate a level of stability that is suitable for practical utilization and are essential to various applications. However, there is a potential for these compounds to be disregarded. The current issue is being tackled by utilizing empirical training data through the implementation of the following methodologies.

The origin of the primary framework for forecasting structure, which is based on empirical evidence, can be traced back to the 1920s. In the decade prior to the current one, the Goldschmidt tolerance factor was a pertinent demonstration of the stability exhibited by perovskites. In contemporary times, novel techniques such as SISSO, gradient tree boosting, and RFs have exhibited superior performance compared to traditional models, leading to a significant enhancement in the prognostic precision for perovskite stability. This improvement has been observed to increase from 74% to over 90%. The Zhuang et al. [30] formulated a hypothesis regarding the potential existence of a substance in a perovskite configuration, either cubic or non-cubic in nature. The result of the prediction produced an average cross-validation error rate of 94%. The application of empirical evidence to forecast stability yields a greater degree of precision and reliability. The rationale behind this is that despite the fact that the convex hull theoretical detachment is a useful metric, it is not completely reliable in determining stability. In order to identify the vast majority of perovskites available in an inorganic crystal model database, it is essential to augment the convex hull detachment by approximately 150 meV/atom, as depicted in **Fig 4**.

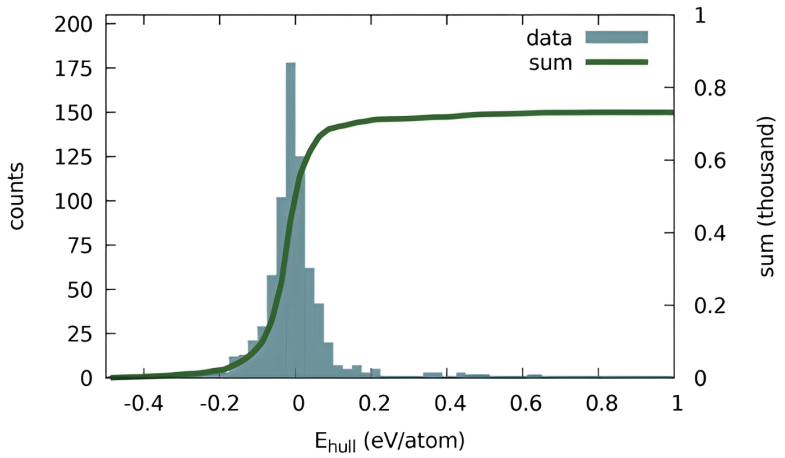

Fig 4. A histogram depicting the distribution of convex hull detachment for perovskites that are encompassed within the inorganic crystal model database

The study conducted involved the performance of 79 calculations using density functional theory, employing the Perdew-Burke-Ernzerhof estimate. The magnitude of the bin width is 25 meV per atom. The AB2C Heusler compounds have been found to possess a significant number of experimentally verified structures. The study conducted by Gao, Lin, and Hu [31] involved the utilization of Random Forests (RFs) and empirical data to develop a prognostic model that determines the probability of producing a comprehensive Heusler compound with a particular composition. The dataset utilized in this model comprised of all compounds exhibiting AB2C stoichiometry, sourced from the alloy phase diagram dataset and the Pearson's crystal dataset. The stability of various novel full-Heusler phases was predicted and experimentally verified by Gu, Ji, Guo, Chen, Yang, and Tan [32] through the utilization of elemental characteristics as fundamental attributes.

IV. CONCLUSION AND FUTURE RESEARCH

This study has conducted an examination of the ML paradigms utilized in the realm of materials science. The research has clarified the essential phases of data processing, including sample construction, data modeling, and model assessment. The present manuscript offers a comprehensive survey of the application of ML methodologies in the domain of materials science. Materials science is a field that utilizes both empirical and computational methods to investigate the structural and functional properties of materials. Subject matter experts (SMEs) with significant experience, spanning multiple years or even decades, employ their knowledge to create innovative materials that demonstrate the intended structure-property relationships. Over the past decade, the domain of ML has undergone substantial growth and has more recently penetrated

various scientific fields such as physics, healthcare, material science, and astronomy. The application of ML has garnered considerable attention in the domains of real estate forecasting, material identification, and investigation of quantum chemistry. This can be attributed to the strong predictive capabilities and cost-effectiveness exhibited in computational operations. However, the utilization of ML methodologies in the domain of materials science is confronted with various obstacles. The field encounters several challenges such as inadequate availability of high-quality data pertaining to materials, challenges in accurately representing material properties, and comparatively lower prediction accuracy when compared to DFT calculations. This paper proposes supplementary approaches that could augment the application of ML methodologies in the domain of materials science.

The expeditious development of material databases holds significant significance for the prospective progress of ML. The efficacy of ML is intrinsically linked to the quality and quantity of data, as it is a methodology that relies on data. The body of experimental records and scientific literature comprises a significant amount of factual information that can be analyzed using ML methods. This information includes, but is not limited to, reaction conditions, synthetic formulations, and molecular properties. The implementation of text mining methodologies enables the efficient acquisition of valuable data that are distributed across diverse articles, journals, and magazines. This procedure has the potential to substantially improve the prevailing material repositories and expedite the establishment of customized databases. Secondly, it is crucial to establish innovative principles for ML. It is anticipated that the utilization of deep learning methodologies and the consequent analogue/manual feature engineering replacement will result in an improved ability to represent raw data in a more efficient manner in the times to come. However, the fundamental principles governing the selection of features and their significance by Deep Neural Networks (DNNs) remain incompletely understood by scholars. The results obtained from deep learning are considered inconclusive and lack a universally applicable theoretical framework. The pursuit of understanding the internal mechanisms of the "black box" not only improves the practicality of ML in the materials science domain but also accelerates the documentation of natural principles, which are yet to be comprehended by humans.

Data Availability

No data was used to support this study.

Conflicts of Interests

The author(s) declare(s) that they have no conflicts of interest.

Funding

No funding was received to assist with the preparation of this manuscript.

Competing Interests

There are no competing interests.

References

[1]. K. Omri, S. Gouadria, M. Madani, S. Mnefgui, N. Alonizan, and F. Alharbi, “Correction to: Doping effects of Ca2+ on the optical and dielectric properties of Ca/ZnO nanopowders materials,” J. Mater. Sci.: Mater. Electron., vol. 34, no. 8, 2023.

[2]. J. H. Kim, T. Y. Shin, S. Yekaterina, and C. K. Park, “Data science approach to find an outlier in the group of cement dispersants,” Constr. Build. Mater., vol. 368, no. 130347, p. 130347, 2023.

[3]. M. R. Raymond, “Assessing performance: Designing, scoring, and validating performance tasksâ□□by Johnson, R. l., penny, J. a., & Gordon, B,” J. Educ. Meas., vol. 46, no. 4, pp. 474–477, 2009.

[4]. S. R. Mohan, S. K. Neogy, A. Seth, N. K. Garg, and S. Mittal, “An optimization model to determine master designs and runs for advertisement printing,” J. Math. Model. Algorithms, vol. 6, no. 2, pp. 259–271, 2007.

[5]. M. Kuanr and P. Mohapatra, “Outranking relations based multi-criteria recommender system for analysis of health risk using multi-objective feature selection approach,” Data Knowl. Eng., vol. 145, no. 102144, p. 102144, 2023.

[6]. M. Büchi, N. Festic, and M. Latzer, “The chilling effects of digital dataveillance: A theoretical model and an empirical research agenda,” Big Data Soc., vol. 9, no. 1, p. 205395172110653, 2022.

[7]. P. R. Punnam, A. Dutta, B. Krishnamurthy, and V. K. Surasani, “Study on utilization of machine learning techniques for geological CO2 sequestration simulations,” Mater. Today, vol. 72, pp. 378–385, 2023.

[8]. K. ur Rehman, J. Li, Y. Pei, and A. Yasin, “A review on machine learning techniques for the assessment of image grading in breast mammogram,” Int. J. Mach. Learn. Cybern., vol. 13, no. 9, pp. 2609–2635, 2022.

[9]. Z. U. Khan, F. Ali, I. Ahmad, M. Hayat, and D. Pi, “iPredCNC: Computational prediction model for cancerlectins and non-cancerlectins using novel cascade features subset selection,” Chemometr. Intell. Lab. Syst., vol. 195, no. 103876, p. 103876, 2019.

[10]. N. Hikichi et al., “Role of sodium cation during aging process in the synthesis of LEV-type zeolite,” Microporous Mesoporous Mater., vol. 284, pp. 82–89, 2019.

[11]. H. Gao et al., “Investigation of contact electrification between 2D MXenes and MoS 2 through density functional theory and triboelectric probes,” Adv. Funct. Mater., vol. 33, no. 15, p. 2213410, 2023.

[12]. T.-Y. Huang, “General framework, mathematical model, current activities and open issues for media adaptation in MPEG-21 DIA: General framework, mathematical model, current activities and open issues for media adaptation in MPEG-21 DIA,” Jisuanji Xuebao (Chin. J. Comput.), vol. 31, no. 7, pp. 1185–1199, 2009.

[13]. E. Tartaglione, A. Bragagnolo, F. Odierna, A. Fiandrotti, and M. Grangetto, “SeReNe: Sensitivity-based regularization of neurons for structured sparsity in neural networks,” IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 12, pp. 7237–7250, 2022.

[14]. G. Wu, “Deep reinforcement learning based multi-layered traffic scheduling scheme in data center networks,” Wirel. Netw., 2022.

[15]. I. Rezaei, S. H. Amirshahi, and A. A. Mahbadi, “Utilizing support vector and kernel ridge regression methods in spectral reconstruction,” Results in Optics, vol. 11, no. 100405, p. 100405, 2023.

[16]. O. Telschow et al., “Preserving the stoichiometry of triple-cation perovskites by carrier-gas-free antisolvent spraying,” J. Mater. Chem. A Mater. Energy Sustain., vol. 10, no. 37, pp. 19743–19749, 2022.
[17]. P. R. Dos Santos, C. T. Pich, D. Back, F. Smiderle, F. Dumas, and S. Moura, “Synthesis, chemical characterization and DNA interaction study of new diclofenac and ibuprofen zinc (II)-nicotinamide ternary complexes as cyclooxygenase inhibitor prototypes,” J. Inorg. Biochem., vol. 206, no. 111046, p. 111046, 2020.
[18]. M. Baumgartner, M. Kropf, L. Haider, S. Veeranki, D. Hayn, and G. Schreier, “ECG classification combining conventional signal analysis, random forests and neural networks - a stacked learning scheme,” in 2021 Computing in Cardiology (CinC), 2021.
[19]. Z. Xia, Q. Zhang, M. Yuan, Z. Liu, and X. Ma, “Structural, magnetic and transport properties of the quaternary Heusler alloy CoFeMnSn,” J. Alloys Compd., vol. 937, no. 168497, p. 168497, 2023.
[20]. S.-P. Korhonen, K. Tuppurainen, A. Asikainen, R. Laatikainen, and M. Peräkylä, “SOMFA on large diverse xenoestrogen dataset: The effect of superposition algorithms and external regression tools,” QSAR Comb. Sci., vol. 26, no. 7, pp. 809–819, 2007.
[21]. S. Dey, A. Das Sharma, and J. Mukhopadhyay, “Effect of oxygen non-stoichiometry and redox phenomena in La/Ba–Sr–Co–Fe–O-based perovskite systems and its heterostructure as applicable in Solid Oxide Cell (SOC) air electrode,” Ceram. Int., 2022.
[22]. W. T. M. Van Gompel et al., “Study on the dynamics of phase formation and degradation of 2D layered hybrid perovskites and low-dimensional hybrids containing mono-functionalized oligothiophene cations,” ChemNanoMat, vol. 7, no. 9, pp. 1013–1019, 2021.
[23]. L. Shen et al., “Three- and two-dimensional mixed metal halide perovskites for high-performance photovoltaics,” Org. Electron., vol. 118, no. 106796, p. 106796, 2023.
[24]. M. Beglari, N. Goudarzi, D. Shahsavani, M. Arab Chamjangali, and Z. Mozafari, “Combination of radial distribution functions as structural descriptors with ligand-receptor interaction information in the QSAR study of some 4-anilinoquinazoline derivatives as potent EGFR inhibitors,” Struct. Chem., vol. 31, no. 4, pp. 1481–1491, 2020.
[25]. X. Chen, J. Qiao, Z. Wang, W. Sun, and K. Sun, “Layered perovskites with exsolved Co-Fe nanoalloy as highly active and stable anodes for direct carbon solid oxide fuel cells,” J. Alloys Compd., vol. 940, no. 168872, p. 168872, 2023.
[26]. A. A. Belik, “Comments on the paper ‘Influence of Ba-doping on the structural and physical properties of Sr2−Ba FeVO6 double perovskites,’” J. Alloys Compd., vol. 940, no. 168932, p. 168932, 2023.
[27]. S. Cheng, Z. Wang, B. Yang, and K. Nakano, “Convolutional neural network-based Lane-change strategy via motion image representation for automated and connected vehicles,” IEEE Trans. Neural Netw. Learn. Syst., vol. PP, 2023.
[28]. R. Gurnani, C. Kuenneth, A. Toland, and R. Ramprasad, “Polymer informatics at scale with multitask graph neural networks,” Chem. Mater., vol. 35, no. 4, pp. 1560–1567, 2023.
[29]. X. Li and Y. Cheng, “Understanding the message passing in graph neural networks via power iteration clustering,” Neural Netw., vol. 140, pp. 130–135, 2021.
[30]. R. Zhuang et al., “Effect of molecular configuration of additives on perovskite crystallization and hot carriers behavior in perovskite solar cells,” Chem. Eng. J., vol. 463, no. 142449, p. 142449, 2023.
[31]. C. Gao, H. Lin, and H. Hu, “Forest-fire-risk prediction based on random forest and backpropagation neural network of Heihe area in Heilongjiang Province, China,” Forests, vol. 14, no. 2, p. 170, 2023.
[32]. Q. Gu, J. Ji, F. Guo, H. Chen, T. Yang, and X. Tan, “First principle study on the electronic, magnetic and phase stability of the full-Heusler compound Fe2CuSi,” SPIN, vol. 11, no. 01, p. 2150001, 2021.