Machine Learning Approches for Evaluating the Properties of Materials

Nanna Ahlmann Ahm
Department of Mechatronics, University of Southern Denmark, Denmark.
ahmann@sdu.dk

Correspondence should be addressed to Nanna Ahlmann Ahm : ahmann@sdu.dk.

Article Info
Journal of Computational Intelligence in Materials Science (https://anapub.co.ke/journals/jcims/jcims.html)
Doi: https://doi.org/10.53759/832X/JCIMS202301007.
Received 08 March 2023; Revised form 08 April 2023; Accepted 25 May 2023.
Available online 10 June 2023.

Abstract – Machine Learning for Materials Science is a primer on the subject that also delves into the specifics of where ML might be applied to materials science research. With a focus on where to collect data and some of the issues when choosing a strategy, this article includes example approaches for ML applied to experiments and modeling, such as the first steps in the procedure for constructing an ML solution for a materials science problem. The lengthy cycles of development, inefficiencies, and higher costs of conventional techniques of material discovery, such as the density functional theory-based and empirical trials and errors approach, make it impossible for materials research to keep up with modern advances. Hence, machine learning is extensively employed in material detection, material design, and material analysis because of its cheap computing cost and fast development cycle, paired with strong data processing and good prediction performance. This article summarizes recent applications of ML algorithms within different material science fields, discussing the advancements that are needed for widespread application, and details the critical operational procedures involved in evaluating the features of materials using ML.

Keywords – Machine learning (ML), Artificial Intelligence (AI), Density Functional Theory (DFT), Artificial Neural Networks (ANN).

I. INTRODUCTION

Machine learning (ML) is the study of how computers learn, specifically how to use information gathered from past experiences to perform better on future tasks. As such, it is considered a subfield of Artificial Intelligence (AI). In ML, algorithms develop a model using sample data in order to produce predictions or judgements without being explicitly programmed. Training data refers to information used to teach a model. Voice technology, computer vision, agribusiness, email screening, and medical supply are just a few examples of the many applications of machine learning algorithms. Not all ML approaches are considered statistical, but a significant part of them are linked to computational statistics that is concerned with providing insights using computers. Computational optimization provides the ML field with novel tools, potential application areas, and multiple theoretical frameworks. Exploratory analysis of data using unsupervised ML approach is the major focus of data mining, which is a related domain of research according to Cazzolato, Rodrigues, Ribeiro, Gutierrez, Traina Jr, and Traina [1]. These learning approaches are based on the assumption that effectual traditional approaches, algorithms, and conclusions will be successful in the future as well.

Inferences like "the sun will rise tomorrow since it has every morning for the last 10,000 days" are examples of the former kind i.e., effectual traditional approaches. For example, "X% of families contain geographically different species with color variations, hence there is a Y% probability that unknown black swans exist." Machine learning algorithms may complete tasks even when not given specific instructions. It is the process of training a computer to do a job based on input data. Computers do not have to learn how to complete routine tasks since they can be provided with algorithms, which detail each action to successfully complete a specific task. It might be challenging for an individual to manually compute the require algorithm for increasingly complex tasks. In real-life, it could be more effective to assist computers to develop their own algorithms compared to having a human programmer specifically explain each step. If an adequate algorithm is not readily accessible for a given job, researchers in the field of machine learning (such as Asnaashari and Krems [2]) will apply a variety of methods to teach computers how to solve the problem. When there are multiple solutions to select from, one approach is to mark as "valid" most to the correct ones. This may be sent back into the computer's algorithm(s) as training data, helping it produce more accurate results in the future. The MNIST dataset of handwritten digits is one example of a dataset used to train a system for a certain purpose.

The experimental trial-and-error approach is the initial paradigm, followed by the chemical and physical laws, computer simulations, and research powered by big data [3]. Among these, the fourth paradigm may completely unite the previous three in terms of theory, computer simulations and experimentations, due to the prevailing advancements technologies such as artificial intelligence and data mining. The study of materials science has led to the development of new approaches that are centred on big data, e.g., Machine Learning. Important and varied material needs have arisen in response to recent advancements in information science, energy, and national security. However, conventional approaches to materials discovery, such as Density Functional Theory (DFT)-based and empirical trials and errors approach, are notoriously time-consuming, expensive, and inefficient, and they struggle to keep up with the rapid advancements in materials science that have occurred in recent decades. In place of time-consuming and expensive methods like DFT calculations or repeated laboratory trials, machine learning offers a promising alternative that may drastically cut development time and expenses.

In 1959, Samuel developed machine learning, which has since found widespread use in fields including as computer vision, economics, bioinformatics, general game playing, and data mining. Now that AI and ML have matured, not only are academics in the traditional AI sector making significant strides forward, but so are specialists in other domains who are applying these methodologies to achieve their own goals. Machine learning was first utilized to identify C60's solubility in materials science in the early 21st century, and has since been put to use in many other areas, including material discovery, property prediction, quantum chemistry research, medication design, and more. Now more than ever, using machine learning in materials science is feasible because of the widespread availability of relevant information and tools for the technique.

This article not only lays out the groundwork for using machine learning to analyze materials' properties, but it also provides a summary of the algorithms' recent applications in a number of well-established areas of materials science and a discussion of the refinements that are needed before they can be widely used. This study aimed to disseminate fundamental machine learning concepts and advocate for their use in the field of materials science. The rest of the article is organized as follows: Section II focuses on the aspect of data processing, discussion data selection and feature engineering. Section III concentrates on modelling, which Section IV discusses model validation. In Section V, a discussion of the application of ML in materials science, and material research process is provided. Lastly, Section VI concludes the research and provides future research for ML in materials science and research.

II. DATA PROCESSING

The advent of big data has resulted in a significant expansion of available knowledge in the materials research field. The effects of big data were broken down into seven categories by Büchi, Festic, and Latzer [4]: value, veracity, visualization, variability, volume, variety, and velocity. Resultantly, these obstacles have hampered the use of data processing in the domain of materials science that is an essential part of machine learning and has an impact on the final model's efficiency. Selecting relevant data and developing appropriate features are the two main components of data processing.

Data selection

Data selection is the process of picking the right data for a project, taking into account criteria including relevance, accuracy, and usability. Those who say "trash in, rubbish out" are not kidding. This essentially indicates that it is irrelevant whether or not a predictive model was constructed if the data used to construct the model was unrepresentative, of poor quality, or otherwise flawed. Success in applying a ML solution is highly dependent on the data utilized in the solution, including its quality, amount, preprocessing, and selection.

Preventing selection bias should be the top priority for any endeavor that expects positive results. When the samples used to create the model are not really representative of the instances the model may be used for in the future, especially with fresh and unknown data, this is known as selection bias. Missing values, worthless values (like NA), outliers, and other anomalies are commonplace in data because of how chaotic it usually is. First, raw data must be pre-processed by means of parsing, cleaning, and transformation before it can be used for modeling and analysis. Common terms for this process include "data munging" and "data wrangling." Data imputation is a common method for filling in gaps caused by missing information; it's conceptually similar to interpolation and serves the same purpose.

Values of features are also occasionally normalized and/or scaled (feature scaling) (normalized). The most common approach to normalizing feature data is to remove the mean across all observations for a specific feature from each observation's value, and then divide by the standard deviation of that feature's observations. While performing machine learning optimization algorithms, feature scaling is used to normalize the range of values for each feature so that no one feature may monopolize the model or the predictions (speed, convergence, etc.). While doing preprocessing, dummy variables may be created to serve as placeholders for qualitative data before being converted to their quantitative counterparts. Changing the values of a color feature (such green, red, and blue) to 1, 2, and 3 is an example. Hence, qualitative traits may now be used in regression analyses.

Mitnitski, Mogilner, MacKnight, and Rockwood [5] should only utilize data from reliable sources to reduce the likelihood of considering irrelevant or irrelevant information. The United States presented the MGP (Materials Genome Project) in 2011 to emphasize the significance of large data sets in driving forward materials research, which in turn actively advocated for the creation of a high quality materials dataset.

High-quality material database

In the computational materials science field, many different databases have already been put to use. These include the Inorganic Crystal Structure, Material Project, Open Quantum Material, Harvard Clean Energy Project, Computational Materials Repository, and AFLOWLIB. Text mining has also been used to extract material-specific literature to supplement pre-existing databases. The Syamsiana, Wibowo, Hakim, Ridzki, and Firjatoellah [6] propose a use for data management in materials engineering, namely in the construction of a ML model informed by the use of data from inefficient experiments. They used experimental data from less-than-effective hydrothermal synthesizing operations to construct a model that predicts when template vanadium selenite crystals would form. As compared to manual analysis, this model performs far better, with an accuracy of up to 89% when predicting the circumstances in which novel organic template inorganic compounds would arise.

Currently, there are four main categories that may be used to describe materials science data: experimental and simulated material properties (physical, chemical, structural, thermodynamic, dynamic, etc.), data from different chemical reactions (such as rate of reaction, and temperature of reaction), datasets from images (such as materials surface images, and electron microscope images), and literature data. These datasets might be either discrete (like text) or continuous (like vectors and tensors) or even a weighted network. Considering data from many databases is made complex due to the idea that data are stored in different databases using different formats. Moreover, the applied machine learning method will dictate the necessary data format. To make use of ML algorithms in the processing of data, it is required to standardize the data on the basis of formats and choose an appropriate data representation. Popular data representations include weighted graphs, fingerprints, SMILES, and Coulomb matrices.

Feature engineering

Feature engineering refers to the generation of meaningful characteristics from unprocessed data. Information from the real world is often unorganized and difficult to parse. The unprocessed data must be converted into a more manageable format before a machine learning algorithm can be applied to it. Data preparation is a step in this direction, and feature engineering is one of its many components. The term "feature" is used to describe a data property that is significant to the issue you want to address using the data and the machine learning model. So, how features are constructed is conditional on the nature of the issue, the information at hand, and the specifics of the machine learning algorithm in use. It would be counterproductive to utilize the same dataset to generate features for two separate challenges. To add to this complexity, several algorithms have specific feature needs. Processes of Feature Engineering are briefly shown in **Table 1**.

Table 1. Feature Engineering Processes

Feature Construction	The process of making use of existing data to build novel features. Understanding the data and the situation at hand is essential for successful feature creation.
Feature Selection	Choosing a selection of characteristics for model training based on their relevance to the situation at hand.
Feature Extraction	Combining and paring down features to make room for new ones that are both more efficient and more broadly applicable. Features may be extracted using several techniques, such as principal component analysis (PCA) and embedding.

To maximize the effectiveness of a machine learning model, feature engineering must be a central element of every use case. Accuracy may be improved by adding features that are both relevant to the issue and suitable for the model. On the other hand, if irrelevant characteristics are input into a data evaluation or systems of ML, the process will provide "Garbage in-Garbage out" results. After the right data has been chosen, the next step is feature engineering, which entails extracting relevant properties of the projected target. Extracting features from raw data for use in algorithms is known as "feature engineering." For machine learning models in general, it may be the limiting factor in how well they function.

Features must be hand-picked in order to use shallow learning techniques (the traditional approach to machine learning). To investigate possible Heusler compounds and their characteristics, for instance, Mitra, Ahmad, Biswas, and Kumar Das [7] turned to machine learning techniques. Experiments were used to settle on 22 characteristics (such as element B's group number, total amount of the A/B radius variation, and p valence electron) that would facilitate the detection of hidden correlations by computers and speed up the investigation. For the purpose of ML model training to foretell undiscoverable HOIP (hybrid organic-inorganic perovskites) for photovoltaics, Patra, Nemade, Mishra, and Satapathy [8] narrowed the feature space from 30 to 14. These 14 features include the factor of tolerance, overall ionic charge, overall sum of p and s orbital radii, and p orbital electrons.

Nevertheless, feature engineering that must be done by hand is not a viable option. The selection of the most representative characteristics for the prediction aim is made harder by the human experience constraints. More time and money are needed for manual feature engineering. A possible future direction for machine learning in materials science is the elimination of the requirement for manually featured engineering due to advances in deep learning.

III. MODELING

A model for material analysis may be constructed with enough data in the right format. The modeling process involves picking suitable algorithms, training using data, and producing reliable forecasts. Separate branches ML integrate reinforcement learning, unsupervised learning, and semi-supervised learning. In the field of supervised learning, sometimes called "learning with a teacher," the training data's matching outputs are explicitly labeled. With unsupervised learning, on the other hand, the matching output of the learning data is rendered un-labeled. As for semi-supervised learning, it is the training data subset that is labelled; in fact, the volume of unlabeled data often much surpasses the overall number of labelled dataset. Instead of instructing a model on how to create the right behaviors, reinforcement learning uses reinforcement signals from the environment to judge the efficacy of the produced responses and fine-tune adaptation tactics.

Moreover, choosing the right ML model is crucial since it has a major impact on the results that may be anticipated. There is a deficiency of a perfect solution that works in every circumstance. Based on the ratio of training data features to labels, ML approaches may be classified as supervised, semisupervised, or unsupervised. To facilitate supervised learning, it is necessary for there to be a one-to-one relationship between input and output data. Computers may learn to predict output values given input values by using supervised models, which allow the computer to discover the link between input and output. In semisupervised learning, the number of input data points often exceeds the number of output data points. There is usually a large portion of incoming data that lacks labels. It is generally accepted that automated training on unlabeled data has a significant impact on model quality. Since the training data lacks labels, unsupervised learning is possible. With unsupervised learning, it is possible to expose the underlying patterning in data. Classification methods using unsupervised learning, such as the K-means algorithm, have found the most widespread use. Of these algorithms, supervised learning is now the most popular one utilized in materials science field.Consequently, in the following, we will only consider supervised learning models.

k-Nearest neighbors

The k-Nearest-neighbours (kNN) [9] algorithm is considered as one of the most foundational widely used machine learning methods. The key concept underlying this method is to identify samples based on how close they are to one another in the feature space. As can be seen in **Fig 1**, when k is equal to 1, the yellow sample is placed in the green category. The sample has k = 3, hence it is red since there are more red triangles than blue ones. If k = 5, then the sample belongs in the "green" category. The k-nearest neighbors technique has potential applications in both regression and classification. In order to classify data, the feature space distance between the training data and the sample must be calculated. In more generic situations, Euclidean distance is always used for accuracy; however, in n-dimensional vector spaces, Minkowski distance may be used instead. In case the position of the samples in the feature spaces are already known, then an explicit training phase is unnecessary. As kNN does not do any training data generalization until the query is received, it is a lazy learning strategy. This indicates that in case the training dataset is massive, the prediction via kNN will be slow and memory-intensive. The efficiency of kNN may also be diminished by an inequitable distribution of training data.

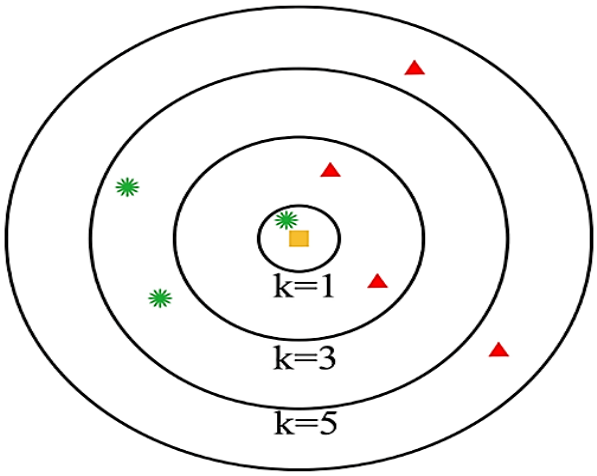


Fig 1. A Representation of the K-Nearest Neighbours Method

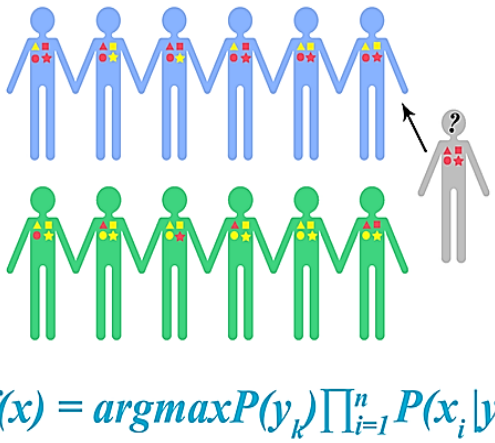


Fig. 2: Computational Function and Naïve Bayes Basic Classifier

The value of k may be chosen arbitrarily. First, cross-validation is employed to identify the modest values of k, which are in correlation to the distribution of samples.

Naïve Bayes

The Naïve Bayes classifier represents a segment of classification algorithms in reference to the Bayes theory and the concept that features are independent to each other. For example, in **Fig 2**, the colour of the features depicted by triangle is

not linked to the features depicted by squares. The samples may be allocated to the class for which the computed probability is the highest. This approach is typically used for the forecasting of biological attributes. Yet, because the traits are constantly connected, it is challenging to meet the premise that they are conditionally independent. In **Fig 2**, we see the central mathematical idea and a straightforward application of the Naive Bayes classifier. The population is split in half, labeled y_1 and y_2. There are four distinguishing characteristics of every individual, and their relative importance is shown by a distinct shade of color (red or yellow). The Naive Bayes approach is used to train a model $f_{(x)}$, and the model's output is the $y_{(k)}$ category, which increased the value of the function. Gray man stands for an unsorted sample, whereas blue and green men stand for two different sorts of samples. Each sample's primary characteristics are represented by one of four geometric shapes, with the shapes' colors standing in for their optional value of characteristics.

Decision tree

The ML predictive model known as a decision tree (DT) consists of nodes and directed edges [10]. Both internal and external nodes are represented here. In **Fig 3**, the central nodes stand for unique feature criteria, while the leaf nodes represent distinct classes. DT has a number of drawbacks, though, including the tree's potential non-robustness. Overfitting might occur if this strategy results in trees that are too complicated. Pruning is a common technique used to speed up and enhance the accuracy of the categorization of fresh data by removing branches that are not dependable and so reducing the complexity of trees and preventing overfitting.

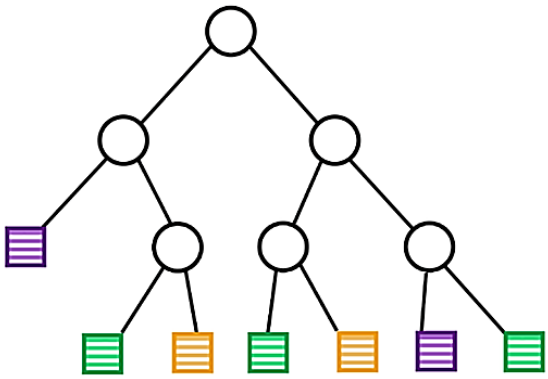


Fig 3. The Decision-Making Process Represented by A Tree Diagram

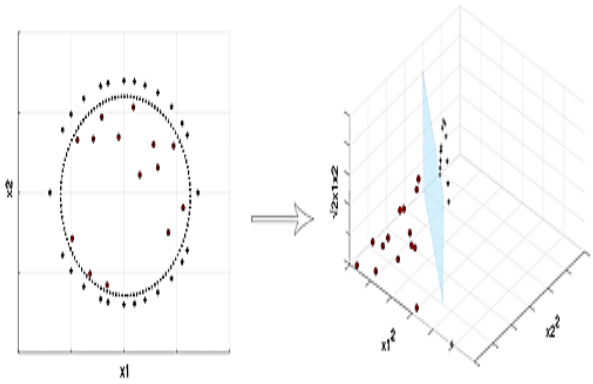


Fig 4. Input data Kernel Transformation Diagram.

Leaf nodes are square while internal nodes are circled. Classes are represented by their respective colors. Random Forest (RF) is a method for classification and regression that combines numerous DTs into a single "forest," first suggested by Asy'Ari et al. in [11]. Each individual tree in the forest is constructed via recursive partitioning. Each DT would provide a verdict whenever a fresh case was introduced. When deciding how to classify an instance, the majority vote would be used, and when using DTs for regression, the average of each would be used. Overfitting may be minimized and large datasets can be handled efficiently with the help of RF.

Kernel methods

To recognize patterns, Deo et al. [12] have developed a set of techniques known as kernel approaches. Popular kernel approaches include the kernel ridge regression (KRR) [13], Gaussian process (GP), and the support vector machine (SVM). As can be seen in **Fig. 4**, these strategies employ a kernel function that is the interior product of the mapping correlation, to increase the dimension of the output space, decrease the computational cost, and enable previously impossible computations. By looking at data in a higher dimensional space, it is easier to see natural groups.

Artificial neural network

Based on biological neural network principles, Artificial Neural Networks (ANN) are a mathematical model for machine learning (ML) and pattern recognition. The model duplicates the approaches for processing complex data in the nervous systems according to the networking topology by first comprehending and then abstracting the structures of the brain system and its response mechanisms. **Fig. 5** depicts the layers of a network, from left to right: input, output, and n hidden $n \geq 1$. Every node has its own unique activation function that it uses to generate output. The weight of each neuronal connection is adjusted during training and then subjected to testing on test data. There is compelling evidence that ANN approaches excel in extracting nonlinear complicated relations from massive datasets. Although ANNs have shown a lot of promise, they still have their drawbacks. Training them often takes a lot of time and a lot of data. ANNs are sometimes referred to as "black boxes" because of how difficult it is to decipher the reasoning behind their decisions. More importantly, overfitting is a problem for ANNs, therefore this approach has to be well thought out before it is used.

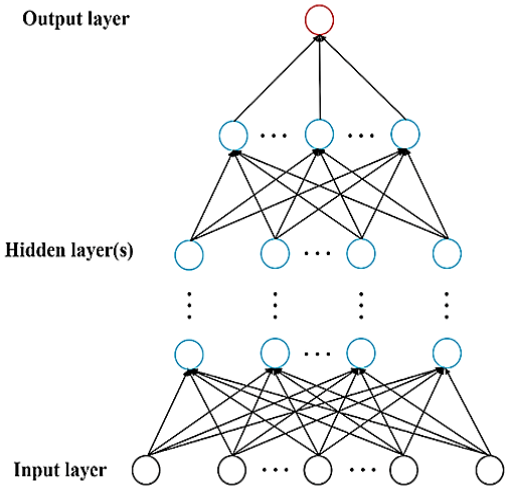


Fig 5. A Prototypical ANN Depiction

In **Fig 5**, the output, input and hidden layers are represented by black, blue, and red circles, respectively. Each dot stands in for a synthetic neuron, and the lines connecting them show how their outputs are fed into the inputs of other neurons.

IV. MODEL VALIDATION

When a mode has fully been trained, it is fully trained on new data that differs from the training data set to see how well it performs. This process is known as model validation. In order to train and validate their models, various ML techniques split the initial set of data into test sets, and training set. Standard validation procedures often use k-fold cross-validation. The K-fold cross-validation alludes to an approach where the initial set of data is randomly sub-divided into k segments, K minus one of those parts are used for training of models in every rounds, and the remaining segment is employed for validation of models. It was with these components in mind that we set out to validate the model, with the final estimate being derived from arithmetic mean of the verification findings. For big data sets, the time investment in building K models for K-fold cross-validation might be prohibitive.

Leave One-Out-Cross Validation (LOOCV) is another frequently used technique for ensuring validity. Similar to K-fold cross-validation, LOOCV considers N models, and uses the averaged classification precision from the previous round of validation for all N models as a measure of classification model. If there are N trials in the original data, then N samples are used for validation and N-minus-one samples are used for training. First, all the datasets in every round are applied to the model's training; therefore, the level of distribution is nearest to the level of distribution of the initial dataset and the results are considered trustworthy; secondly, LOOCV is faster than K-fold cross validation. It is possible to repeat the experiment because there are no uncontrollable variables that may alter the results. The high computational cost of LOOCV is similar to that of K-fold cross validation. It might be challenging to put LOOCV into practice in case the overall number of the initial data sample is enormous unless the computations can be parallelized to shorten the calculation time.

Some more techniques used to check the model's accuracy are the RLT cross validation and the bootstrap cross-validation. RLT cross-validation, in contrast to LOOCV, splits the dataset in half for the validation phase. As a result, there is a significant reduction in the amount of computing complexity required. Although it may seem obvious that the test set should consist of as many observations as possible, in fact, this is not the case. It is possible to reduce the likelihood of making a mistake when extrapolating results by using the multiple sampling techniques known as bootstrap cross-validation. K-fold cross-variance validation's is effectively reduced with this technique, but the computing cost rises as a result.

V. APPLICATIONS

Machine learning outperforms computer simulation in its ability to recognize patterns in high-dimensional data, extract meaningful insights rapidly, and unearth previously unknown rules. As predicting material characteristics often necessitates computationally intensive theoretical computations, this method is ideally suited for material discovery and may speed up the process. Specifically, we will discuss three of the most prominent uses of machine learning in materials science at current time.

Material property analysis

Degradation detection

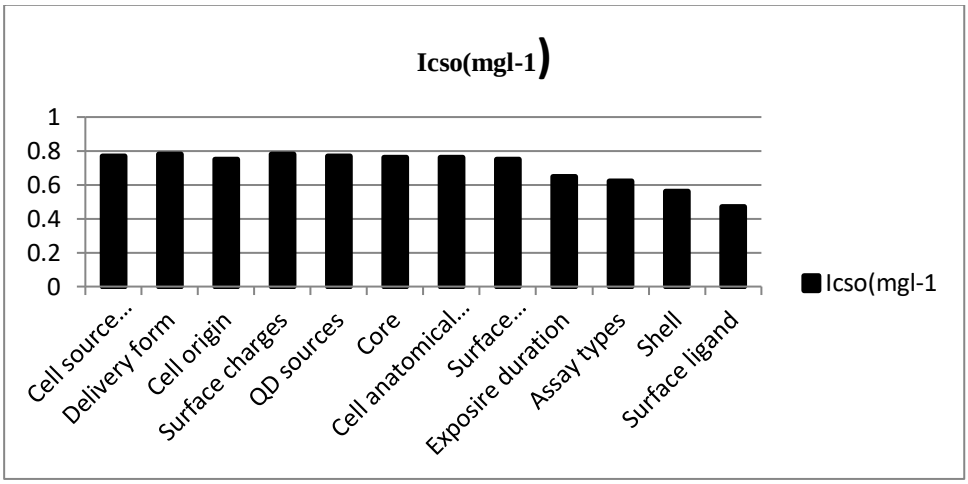
When it comes to material analysis, such as the identification of corrosion in metals or asphalt pavement cracks or the measurement of concretes strength, ML is more convenient and accurate compared to the judgement of humans. With a focus on the link between the alloy's characteristics and its manufacturing and composition process factors, Dupuis et al. [14] investigated the use of machine learning approaches (such as predictive and selective modelling) to predict the steel's

fatigue strength. The tempering temperature was determined to be the most influential of 25 parameters examined in relation to fatigue strength. There were four distinct phases to this procedure. To begin, raw data were pre-processed using domain expertise. Secondly, we used ranking-based feature selection techniques to zero in on the most relevant details. Predictions of fatigue strength were then made using machine learning techniques. We then used LOOCV to assess the models' precision. Several machine learning techniques, including ANNs, SVMs, and linear regression, achieved great prediction accuracy, as shown by R2 values >.98 and error rates 4%, as shown in the findings.

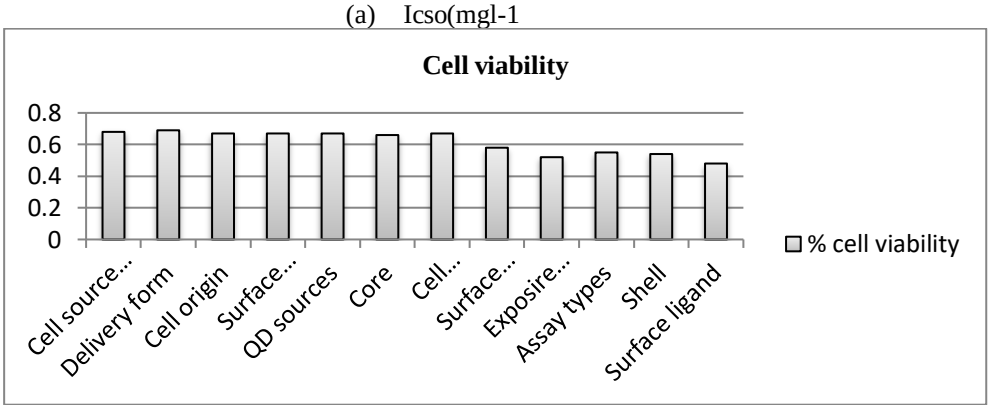
Zhang presented a CNN-based efficient architecture with pixel-level precision. Automatic pavement fracture detection with accuracy of 90.13% was achieved using 3D photos of an asphalt surface. Author's model proposes a useful approach for checking the condition of railroad tracks. We used a fully convolutional network as the foundation for our model. The material classification process used four convolutional layers, whereas the fastener detection process utilized five. Scientists gathered 203,287 images of the route over 85 miles by using an artificially lit automobile, which they then analyzed using specialized software. Using the five-part data set, only 20% of the photos were utilized for testing, while 80% were used for training. A total of 50,000 patches were randomly selected from each class for each data segmentation. Consequently, 2 million patches were employed in training every model.

Nanomaterials analysis

In the study of nanomaterials, machine learning has become more significant as AI has progressed. To investigate C60 solubility, the use of machine learning was suggested as early as 1993. Toxicological predictions of nanomaterials, the identification of novel non-toxic nanoparticles, the creation of single/multi-structure property relations of nanoparticles, the assessment of quantum-mechanical molecular system observables, the assessment of nano-materials chemical reactions, and the solutions of kinetic models are just some of the various applications of ML in the field. The chemical toxicity of quantum dots (QDs) was investigated by Soni, Tripathi, Bundela, Khiriya, and Khare [15], who effectively used meta-analysis. To evaluate the data, Iqbal, Onyelowe, and Jalal [16] employed the random forest regression model after using data mining to generate important information on toxicity of GQs from 307 published researches. Their findings showed that the toxicity of QDs was proportional to factors like their diameter, test type, and duration of exposure, as well as surface features (such as shell, ligand, and surface modification). On display in **Fig. 6** are the salient features of 12 species that have a major impact on QD toxicity.



(a) Icso(mgl-1



(b) Percentage cell viability

Fig 6 (a) and (b). Quantum Dots Containing Cadmium: A Meta-Analysis of Their Biological Toxicity

Quantum dots (QDs) toxicity data was collected through data mining, and random forest was utilized to determine what characteristics of the QD data were most important for modeling QD toxicity.

Molecular property prediction

Predicting molecular properties using high-throughput density functional calculations is an extremely time-consuming process. Machine learning, on the other hand, is an achievable strategy for the rapid forecasting of characteristics or structures of materials, compounds, and molecules, and it can attain higher levels of accuracies. ElemNet model is a DNN-based framework for predicting material qualities given just the components that make up the target material. It automatically extracts the similarities and interactions between elements and their physics and chemistry, allowing for quick and accurate predictions. **Fig 7** depicts the ElemNet architecture. A convolutional neural network (CNN) is at the heart of Chemception, a model that, like ElemNet, translates raw dataset of chemicals into two-dimensional images to foretell an activity, solvation, and toxicity characteristics. Moreover, Li et al. [17] presented a broad machine learning approach for predicting inorganic material characteristics. The model may be used to make predictions about many different features, including glass-forming ability, and band gap energy in both crystalline and amorphous materials.

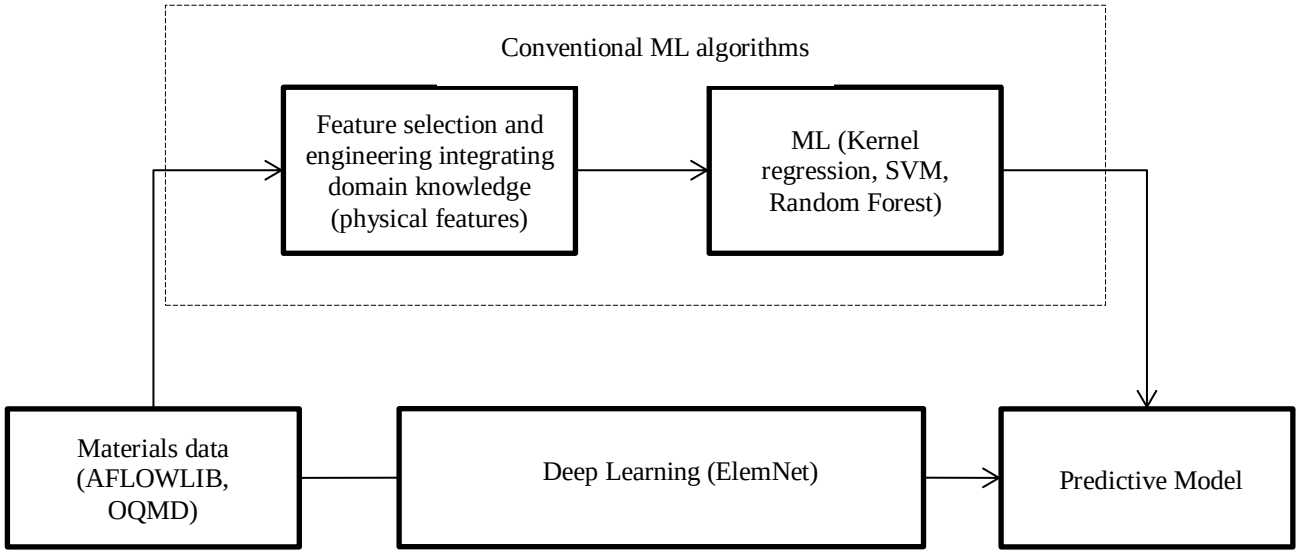


Fig 7. Deep Neural Networks Used to Predict Material Properties

Discovering new materials

Structure-oriented design

In order to infer a material's chemical make-up from its structure, several different classification and regression techniques may be used. The perovskite crystal structure is significant in a wide variety of contexts. Six stable lead-free HOIPs were predicted by Kim, Vasileiadou, Spanopoulos, Kanatzidis, and Tu [18] using a variety of regression algorithms (multilayer perceptron regression, gaussian process regression, kernel ridge regression (KRR), gradient boosting regression, support vector regression, and DT regression) ($C_2H_5OInBr_3$, $NH_3NH_2InBr_3$, $C_2H_6NInBr_3$, $C_2H_5OSnBr_3$, $C_2H_6NSnBr_3$, and NH_4InBr_3). In order to anticipate the electrical band gap of dual perovskite and to select stabilized perovskite users, Saeedi, Bouraghi, Seifpanahi, and Ghazisaeedi [19] established a systematic feature engineering approach and effectual ML model.

Element-oriented design

Predicting novel compounds, including their structural orientations provided the composition of an input is a useful use of machine learning. Approximately 209 novel ternary oxides were discovered with the use of a probabilistic model built on an experimental crystal structure database. Moreover, Stepinski and Dmowska [20] built a machine learning model for forecasting the thermodynamic stability of arbitrary mixtures using hundreds of DFT computational findings. Two predicting formation models of energy and a vast database of DFT calculation results were created by the Researchers (1 ML and 1 heuristic model). The two models were then used to search for 4500 new stable materials among 1.5 million potential ternary compositions having the most probable findings rated using a combination of the two models. Comparable research on binary chemicals also made use of machine learning.

A regression approach was therefore utilized to project the temperature of melting of 44 AB sub-octet elements by mining 16-attribute combinations of component atoms in every binary compound. The first step in this process was to utilize an unsupervised learning approach to categorize about 67 octet elements into different categories based on their

crystalline structures. According to the aforementioned research, machine learning offers both excellent accuracy and great promise for the discovery of novel chemicals.

Recent study has presented an objective-reinforcement generative adversary network-based model for the reward-based generation of novel organic compounds with predefined chemical characteristics and physical reactions. The framework integrates generators and discriminators, with the former capturing data distribution and the latter comparing the generated molecular composition with known molecular components to determine whether the generated components are possible. The discriminator's error probability is optimized via training the generator to produce erroneous results. In order to improve the discriminator's ability to tell the difference between actual and bogus data, you may repeat this procedure. As a result, a reinforcement learning network based on a "reward mechanism" may be utilized to create chemical compounds with novel physical or biological properties. An overview of the structure of this generative adversarial network with an emphasis on objectives is shown in **Fig 8**.

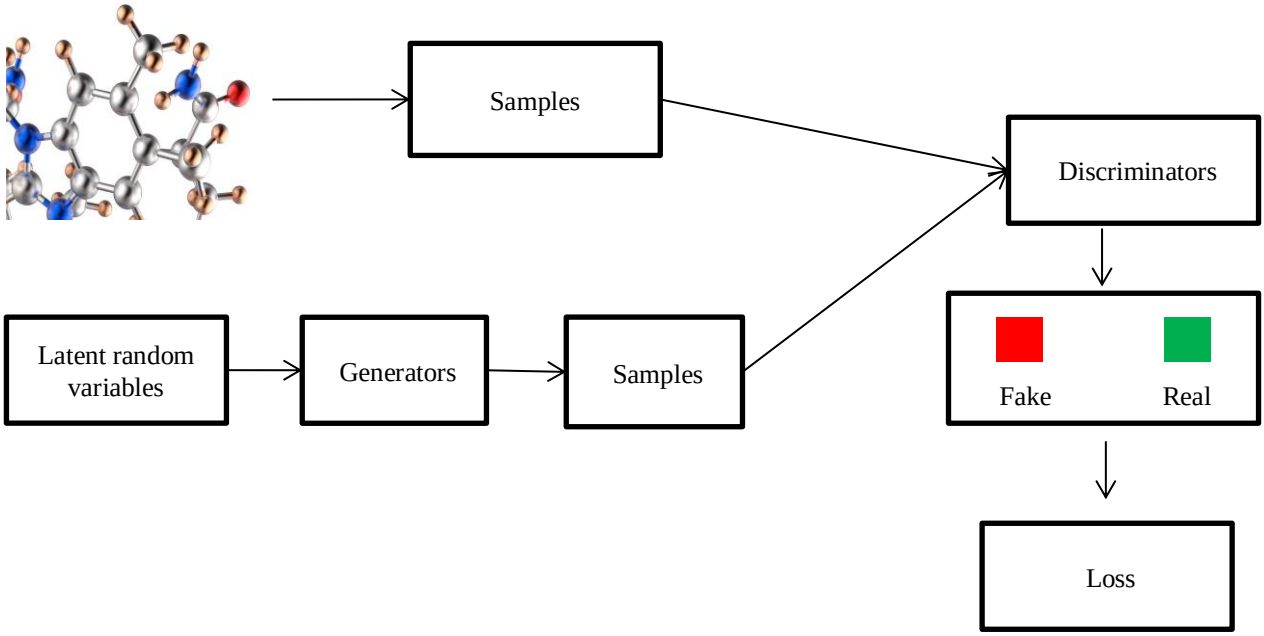


Fig 8. Model on Objective-Reinforcement Generative Adversary Networks

VI. CONCLUSION AND FUTURE RESEARCH

Due to its robust prediction performance and very cheap processing cost, machine learning has found widespread use in the forecasting of features, the identification of novel nanomaterials, and the study of quantum chemistry. Nonetheless, numerous challenges remain in using machine learning in the materials science field. For instance, there is a deficiency in the quantity of high-quality data pertaining to materials, material characteristics are tricky to completely express, and accuracy of predictions is minimal compared to that of the DFT computations. Here, we suggest new avenues that might help advance the use of ML in materials science and research.

First, it is crucial for machine learning's long-term health that we speed up the process of building a materials database. As machine learning is data-driven, the quality and amount of the data available are crucial to its success. Machine learning may be used to a wealth of data found in the scientific literature and experimental records, including information on molecular characteristics, reaction circumstances, and synthetic formulations. This relevant data, dispersed among articles, journals, and magazines, may be swiftly gathered via text mining, which will significantly improve the current material dataset and permit the establishment of specialize datasets. Second, it is crucial to develop new guidelines for machine learning. A more accurate representation of raw data is on the horizon, thanks to the rise of ML and the decline of human involvement in feature engineering. Experts, however, remain confused about how DNNs decide which characteristics to use and what those traits really signify. Due to this, deep learning cannot provide a generally applicable theory and its findings are not as persuasive. Improving the generalisability of ML in material science is merely one benefit of delving into the "black box." Such an investigation may also lead to the discovery of hitherto unknown laws of nature.

Finally, machine learning may find important use in quantum chemistry. Several challenges in quantum chemistry can be solved by machine learning due to its tremendous data-processing capacity. The accuracy of predictions of models could be more enhanced by integrating DFT with ML. This has the potential to become an invaluable resource for the prediction of intricate molecular characteristics and architectures, the study of quantum multi-body systems, and the identification of novel materials. There are still jobs for which machine learning, despite significant progress, cannot perform to expectations because of a lack of appropriate data. Thus, there are cases when a more precise model is required,

one that has been learned using a relatively small but precise dataset. It has been shown that a deep learning model trained with just 4000 examples achieves satisfactory performance. In addition, the previously mentioned approach of training models using failure dataset that was obtained from failed trials could be fundamental in such instances.

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.

Ethics Approval and Consent to Participate
The research has consent for Ethical Approval and Consent to participate.

Competing Interests
There are no competing interests.

References

[1]. M. T. Cazzolato, L. S. Rodrigues, M. X. Ribeiro, M. A. Gutierrez, C. Traina Jr, and A. J. M. Traina, "Sketch+ for visual and correlation-based Exploratory Data Analysis: A case study with COVID-19 databases," jidm, vol. 13, no. 2, 2022.
[2]. K. Asnaashari and R. V. Krems, "Gradient domain machine learning with composite kernels: improving the accuracy of PES and force fields for large molecules," Mach. Learn. Sci. Technol., vol. 3, no. 1, p. 015005, 2022.
[3]. S. R. Heller, K. Scott, and D. W. Bigwood, "The need for data evaluation of physical and chemical properties of pesticides: the ARS pesticide properties database," J. Chem. Inf. Model., vol. 29, no. 3, pp. 159–162, 1989.
[4]. 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.
[5]. A. Mitnitski, A. Mogilner, C. MacKnight, and K. Rockwood, "Data integration and knowledge discovery in biomedical databases. Reliable information from unreliable sources," Data Sci. J., vol. 2, pp. 25–34, 2003.
[6]. I. N. Syamsiana, S. S. Wibowo, M. F. Hakim, I. Ridzki, and R. Firjatoellah, "Energy Database Management System (EDBMS)-based data acquisition audit for electricity savings analysis," IOP Conf. Ser. Mater. Sci. Eng., vol. 1073, no. 1, p. 012036, 2021.
[7]. S. Mitra, A. Ahmad, S. Biswas, and A. Kumar Das, "A machine learning approach to predict the structural and magnetic properties of Heusler alloy families," Comput. Mater. Sci., vol. 216, no. 111836, p. 111836, 2023.
[8]. K. Patra, B. Nemade, D. P. Mishra, and P. P. Satapathy, "Cued-click point graphical password using circular tolerance to increase password space and persuasive features," Procedia Comput. Sci., vol. 79, pp. 561–568, 2016.
[9]. L. Yang, Z. Liu, and Q. Kong, "Power consumption prediction with K-nearest-neighbours and XGBoost algorithm," Int. J. Wirel. Mob. Comput., vol. 15, no. 4, p. 374, 2018.
[10]. Decision Tree Writing Group, "Clinical response Decision Tree for the Mountain Gorilla (Gorilla beringeii) as a model for great apes," Am. J. Primatol., vol. 68, no. 9, pp. 909–927, 2006.
[11]. R. Asy'Ari et al., "High heterogeneity LULC classification in Ujung Kulon National Park, Indonesia: A study testing 11 indices, Random Forest, sentinel-2 MSI, and GEE-based cloud computing," CELEBES Agricultural, vol. 3, no. 2, pp. 82–99, 2023.
[12]. R. C. Deo et al., "Cloud cover bias correction in numerical weather models for solar energy monitoring and forecasting systems with kernel ridge regression," Renew. Energy, vol. 203, pp. 113–130, 2023.
[13]. R. Singh and S. Vijaykumar, "Kernel Ridge Regression Inference," arXiv [math.ST], 2023.
[14]. N. Dupuis et al., "AlGaInAs selective area growth by LP-MOVPE: experimental characterisation and predictive modelling," IEE Proc., vol. 153, no. 6, pp. 276–279, 2006.
[15]. A. Soni, G. K. Tripathi, P. Bundela, P. K. Khiriya, and P. S. Khare, "Synthesis and characterization of assorted pH CdSe quantum dots by solvo-thermal method to determine its dye-degradation application," Opt. Quantum Electron., vol. 55, no. 3, 2023.
[16]. M. Iqbal, K. C. Onyelowe, and F. E. Jalal, "Smart computing models of California bearing ratio, unconfined compressive strength, and resistance value of activated ash-modified soft clay soil with adaptive neuro-fuzzy inference system and ensemble random forest regression techniques," Multiscale Multidiscip. Model. Exp. Des., vol. 4, no. 3, pp. 207–225, 2021.
[17]. G. Li et al., "Obstructed surface states as the descriptor for predicting catalytic active sites in inorganic crystalline materials," Adv. Mater., vol. 34, no. 26, p. e2201328, 2022.
[18]. D. Kim, E. S. Vasileiadou, I. Spanopoulos, M. G. Kanatzidis, and Q. Tu, "Abnormal in-plane thermomechanical behavior of two-dimensional hybrid organic-inorganic perovskites," ACS Appl. Mater. Interfaces, vol. 15, no. 6, pp. 7919–7927, 2023.
[19]. S. Saeedi, H. Bouraghi, M.-S. Seifpanahi, and M. Ghazisaeedi, "Application of digital games for speech therapy in children: A systematic review of features and challenges," J. Healthc. Eng., vol. 2022, p. 4814945, 2022.
[20]. T. F. Stepinski and A. Dmowska, "Machine-learning models for spatially-explicit forecasting of future racial segregation in US cities," Mach. Learn. Appl., vol. 9, no. 100359, p. 100359, 2022.