A Review on Background and Applications of Machine Learning in Materials Research

[1]Robert Ahmed and [2]Christna Ahler
[1,2]University of Akureyri, 600 Akureyri, Iceland.
[1]ahmedrob@hotmail.com

Correspondence should be addressed to Robert Ahmed : ahmedrob@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/JCIMS202301008.
Received 15 March 2023; Revised form 20 April 2023; Accepted 15 June 2023.
Available online 30 June 2023.

Abstract – In recent decades, Artificial Intelligence (AI) has garnered considerable interest owing to its potential to facilitate greater levels of automation and speed up overall output. There has been a significant increase in the quantity of training data sets, processing capacity, and deep learning techniques that are all favorable to the widespread use of AI in fields like material science. Attempting to learn anything new by trial and error is a slow and ineffective approach. Therefore, AI, and particularly machine learning, may hasten the process by gleaning rules from information and constructing predictive models. In traditional computational chemistry, human scientists give the formulae, and the computer just crunches the numbers. In this article, we take a look back at the ways in which artificial intelligence has been put to use in the creation of new materials, such as in their design, performance prediction, and synthesis. In these programs, an emphasis is placed on the specifics of AI methodology implementation and the benefits gained over more traditional approaches. The last section elaborates, from both an algorithmic and an infrastructural perspective, where AI is headed in the future.

Keywords – Machine Learning, Materials Research, Materials Engineering, K-Nearest Neighbor, Artificial Neural Networks.

I. INTRODUCTION

The discovery of new nano-materials will play a vital role in fostering the development of human. Following hundreds of years of research and development, materials science has amassed a wealth of information. As a result of our limited mental capacity, it may be challenging for humans to keep up with the daily deluge of new information. It is only possible to examine a fraction of the data that exists in a certain sector. Today's material studies rely heavily on the "trial-and-error technique," which entails conducting several tests under the guidance of experts and supplementing them with a few computer simulation calculations. This approach requires a significant investment of time, money, and resources. Many important data sets remain unnoticed or underutilized in the repository. Consequently, it is important to discover a new research approach in order to speed up the development of novel materials. The advent of AI heralds a new era in the progress of the physical sciences.

For almost 60 years, AI has progressed from the basic perceptron to the sophisticated multilayer neural networks, demonstrating a main algorithm framework and a robust hardware base. Highly developed AI systems have already bested human world champions in a variety of games and activities, including chess, go, and quiz shows, among others. The field of materials science is paying close attention to AI because of its superior data mining capabilities, as can be seen in lines. Turing Prize laureate James Gray presented his idea for "the fourth paradigm of science" at the NRC-CSTB conference in 2007 [1]. Using big data and AI, this data-heavy field condenses vast amounts of information into novel hypotheses that may direct future research. Several challenges in the field of materials science are analogous to the kinds of scales and nonlinearities that are amenable to this approach. Materials informatics is an interdisciplinary field that combines materials science with artificial intelligence techniques to better assist scientists in determining the hidden relationship between variables, making accurate predictions about material properties, determining the best chemical synthesis path, adjusting process parameters, and improving upon currently used material characterization techniques.

In the pharmaceutical sector, materials characterization methods are employed to establish medications' essential properties. Material characterisation data especially that of powders and liquids, is crucial to the manufacturing and effectiveness of medicinal products. New methods of characterizing nano and microparticles are being developed with help from materials science and engineering. **Table 1** displays a selection of the materials characterisation methods used in the pharmaceutical business. It is common knowledge among materials engineers to use tools like Nuclear Magnetic

Resonance (NMR), Raman microscopy (RM) and X-ray diffraction (XRD). They are widely used for research into the defects, surface uniformity, percentage crystallinity, and amorphous content of powders employed in the pharmaceutical sector.

Table 1. Materials Characterization Techniques

Dynamic light scattering and photon correlation spectroscopy	They are used in the pharmaceutical business for analyzing particle sizes (nm and pm). Drug performance, stability, and dissolving time are all influenced by particle size. Particle property manipulation is also crucial for maximizing the efficiency of reactions between materials and catalysts during manufacturing.
Mercury intrusion	A tablet's porosity may be determined using this. Friability (the propensity to break part), solidity, and dissolving behavior are all affected by a tablet's porosity.
Gas density pycnometry and energy density analysers	Tablets' densities are measured using these in order to ensure they are high enough.
Differential scanning calorimetry (DSC), and Thermogravimetric analysis (TGA)	TGA and DSC are well-known methods for thermal characterisation of materials. Similarly, the pharmaceutical business makes use of these methods.

Machine learning (ML) is a fast growing discipline of artificial intelligence (AI) that has the greatest promise for use in the study of materials science. Section II provides groundwork of ML fundamentals for the subsequent discussion of AI's applications in the field of materials research. Section III focuses on a critical survey of AI application for material science and engineering. The last Section IV draws a conclusion to the paper as well as directions for future research.

II. BACKGROUND ANALYSIS OF MACHINE LEARNING

Machine learning (ML) is the process by which a computer is taught to analyze a body of data in order to discover the rules or other hidden knowledge that underlies it. In particular, ML consists of the following four phases: data representation, model optimization, method selection, and data collection. ML's many subcategories and algorithmic variants are shown in **Fig 1**.

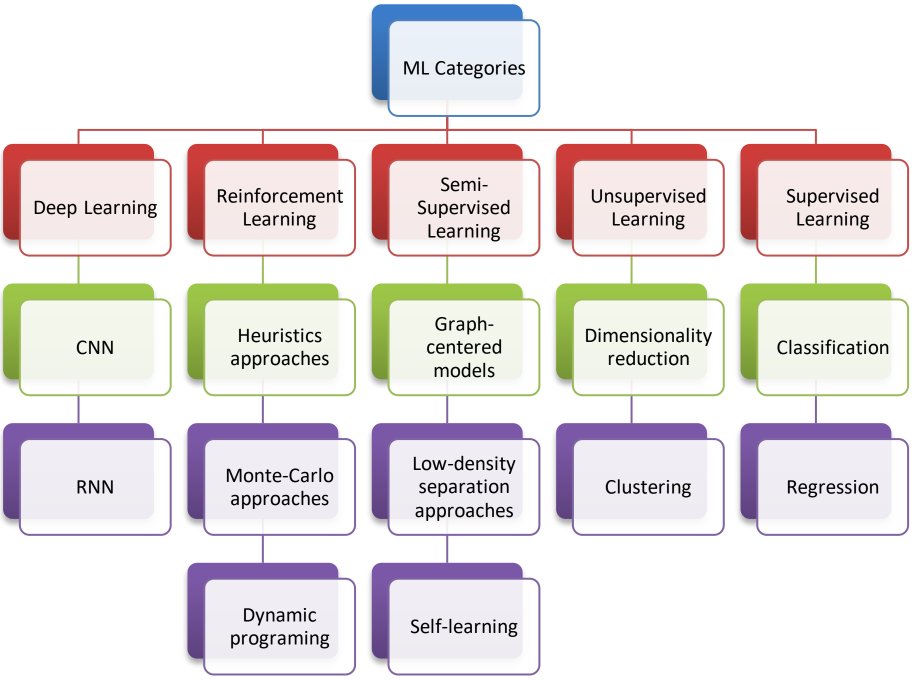


Fig 1. Types and Methods of Machine Learning

Data Collection

As a type of data-driven algorithms, ML may draw on information gleaned through simulations (including DFT and MD), experiments, and online databases. Certain materials' structures and physical characteristics are included in the database. Because of constraints imposed by the environment and experimental methods, there is a great deal of incomplete, duplicated, and inconsistent data in the area of materials. To this end, data cleaning—the process of discovering and fixing various problems with raw data—becomes a rather vital one. When a value for an attribute is absent, it is often filled in using the minimum, average, and other statistical value for that property.

The process of sorting by integrating records, and feature values with equal value is a fundamental concept for removing duplicate entries with repeated values. In this context, relevant algorithms integrate the queue algorithm priority, the sorted neighbourhood approach, and others. Using this strategy, Hyde and Andersson [2] were able to combine data from the Inorganic Crystal Structure and the Materials Project Databases to better understand perovskite structures. Specific programs may be built to assess whether or not the data fulfill the criteria if the values are inconsistent, taking into account the acceptable value range and the mutual connection of each variable. Out-of-bounds information or data with competing qualities will be removed. The cleaned data may then be represented in a meaningful way.

Data Representation

Data is an accumulation of discrete figures, which transmit data in the quest of skills and knowledge. The values could define the qualities, statistics, facts, quantities, or other relevant units of knowledge, or they could be characteristic sequences, which may be defined in multiple ways. A datum or piece of data represents a single numeric element within a dataset. The structural elements such as tables are employed to organize data, which could be reused at values in other contexts. The data may be treated as independent variables in the calculation. Information may be interpreted as either ephemeral concepts or tangible quantities. Data is utilized widely in different fields of human organization, including but not limited to politics, economics, and natural science. Data sets contain things like census information, unemployment statistics, rates of literacy, and price indexes. Data, in this sense, is the raw numbers and facts, which can be manipulated to produce insights or data.

The analysis, inquiry, observations and measurement are some of the approaches, which can be used to provide meaningful data that is then considered as characters or numbers, which can be processed further. Information gathered in the field is data gathered in a natural setting. The term "experimental data" [3] refers to any and all information gathered in the course of a well supervised scientific experiment. Calculation, visualization, presentation, reasoning, discussion, and other forms of post-analysis are employed to evaluate data. Raw data (or unprocessed data) is generally cleansed before analysis: Errors caused by outliers are eliminated, while mistakes caused by faulty instruments or data input are fixed.

In this context, "data" refers to any discrete pieces of information that may be utilized as inputs to a computation, as outputs from a logic process, or as topics for conversation. Statistics are only one kind of data that may vary from conceptualizations to actual measurements. Information is any collection of material that has a discernible theme and is presented in a meaningful setting. Data insights or intelligence are therefore a description of such contextually related information. Knowledge is the accumulated store of ideas and wisdom that comes from turning raw facts into usable information over time. "The new oil of the digital economy" is data, as it has been said. In simple terms, data is any information or knowledge that has been represented or codified in a way that makes it more usable or more amenable to processing.

Big data, the term used to describe massive data sets, has emerged as a result of technological advances in computing. Some datasets may exceed a petabyte in size. Using such big sets of data is challenging, if not ineffective, using traditional methods of data analysis and computation. (Theoretically, unlimited data would result in endless information, making it impossible to draw any conclusions or use any intelligence from it.) In response, data science, a relatively young profession, employs machine learning (and other artificial intelligence (AI) approaches) to enable effective applications of analytical methods to huge data.

To "represent" data is to transform information into a form that can be processed by a computer program. The information we get is often numerical, but it may not be suitable for the algorithm. In the same way that listing equations or plotting important figures is helpful for solving mathematical issues, so too can listing useful information. In addition to needing the right kind of input data, ML algorithms cannot improve their learning abilities without it. The better the model does when we pick a more suitable representation. Binary coding is one way to convey details about a structure's physical characteristics. In a paper published in Nature, Lim, Leow, Pham, and Tan [4] presented a robot that could do organic synthesis. Robots can evaluate the level of reaction of catalytic combinations and apply SVM (support vector machine) models to predict unidentified chemical reactions thanks to binary encoding of the chemical input.

Algorithm Selection

Unsupervised learning, which integrates clustering, and supervised learning, which integrates regression and classification, are the two main categories of ML. Recent advances in materials automation have led to the development of learning methods that need interaction with their surroundings; these include reinforcement learning and active learning. The decision tree (DT) symbolic regression, artificial neural networks (ANNs), and k-nearest neighbor (KNN) algorithm are now the most often used algorithms.

KNN is an efficient and straightforward technique for classification and regression. The method takes a new datum and a training dataset and identifies the k closest entries to the new datum in the dataset, then assigns the new datum to the category in which it occurs most often. The method is made up of three parts: choosing k, measuring distance, and applying a categorization criterion. Inversely, when k increases, the complexity of the model tend to decrease; the approximation error decreases, and the estimate error increases. Two points' similarity may be determined differently depending on the distance used to compare them. KNN has a variety of distance metric options including the Euclidean distance, the Manhattan distance, and others. Empirical error reduction is the goal of KNN; hence it often chooses majority voting as a criterion of classification.

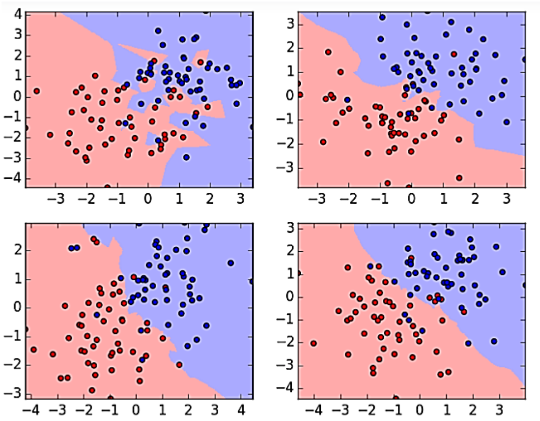
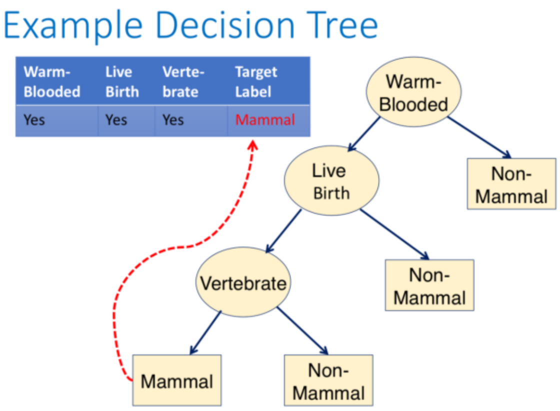


Fig 2. Plots Showing Identical Data Sets with Different K Values


Example Decision Tree

Warm-Blooded	Live Birth	Vertebrate	Target Label
Yes	Yes	Yes	Mammal

Fig. 3: Decision Tree Example

We need to know how the value of K affects the algorithm before we can evaluate what factors should be taken into account. Two plots of the same data set are shown in **Fig 2**; the leftmost plot has a K-value of 1, while the rightmost plot has the greatest possible K-value. If we take a close look, we can see that as K becomes larger, the border of the classification method gets smoother. Specifically, the smoothness of the border is proportional to the Magnitude of K. This explains why a K value of 1 will cause the training model to overfit the data, while a high K value will cause the model to underfit the data. Checking the validation error for a variety of K values and selecting the one with the smallest error will help us determine the best K number to use.

A decision tree is a simple and powerful machine learning (ML) algorithm. A decision tree is a representation of a classifier that accepts input consisting of several attribute values and produces an outcome. It is possible for both the input and output values to be continuous or discrete. Boolean classification is used when there are only two potential outcomes and the inputs are discrete. A decision tree will go through a series of tests and then provide its verdict. Each node in a decision tree is a potential value for one of the attributes, and the sub-branches falling off are the tests of the values of the attribute. All of the function's returned values are its leaf nodes. **Fig 3** is a decision tree illustration. Decision trees are a method for classifying data in which the data is divided into subsets that each correspond to a single class. To do this, the input space is segmented into uncontaminated areas.

There are several traditional AI algorithms, but one of the most popular is genetic programming-based symbolic regression (GPSR). Since the functional link between variables is not assumed, this method deviates from the standard numerical regression. Instead, chromosomal evolution in each putative function is what yields the functional form. The chromosomes have both internal and external nodes, the former containing symbols for operations and the latter containing variables and constants. Chromosomes may be walked through using the depth-first search strategy in order to find the associated function. The evaluation function is the difference in errors between the experiment data, and the data integrated by functions. Candidates with the lowest inaccuracy and highest adaptability may be given preference when it comes to creating offspring. Many chromosomes go through processes of mutation and inheritance, iterating until the optimal function and parameter set are discovered. The grapheme magic angle, the regular hydrogen viscosity, and quests for perovskite stability descriptors are just a few examples of the fields where GPSR is well suited to explore materials with limited previous information and ambiguous relationships between related variables.

The creation of artificial neural networks is motivated by the theory that thought occurs largely as a result of electrical activity within the brain cell network known as neurons. The building blocks of a neural network are nodes linked by directional connections. Every connection between nodes has a weight, expressed as a number, which specifies the magnitude and direction of activation propagation. Two primary methods exist for establishing links between nodes in a network. A feed-forward network only has connections going forward between nodes. In computer science, a recurrent network is one in which the outputs are used as inputs to another network. Just like the hidden, output and input layers, the most popular network also has additional layers. The goal of the learning process is to optimize the output accuracy by determining the best settings for the parameters. The model is solidified after training and strategy testing.

Not just the aforementioned, but also kernel approaches, generative adversarial networks, random forests, and convolutional neural networks are more effective ML algorithms. Certain hyperparameters of the chosen method must be calculated by hand or by heuristic techniques. More and more studies are being conducted on autonomous ML, which tries to simplify the process of using ML techniques.

Model Optimization

The model with high degree polynomials could fit the learning data well, but if this degree is increasingly higher, the model might overfit and report worse performance on data validation. Cross-validation and regularization are two methods for selecting the polynomial degree that together aim to reduce the weighted summation of the model’s complexity and the empirical loss. While looking for a model, loss functions are often employed to find one with the lowest potential error rate. To quantify this gap between true and anticipated values, we develop the loss function. An optimal hypothesis could be identified by minimizing a loss function. In case the samples employed for validation and training in cross-validation are not actually representative of the entire population, then the results will not be accurate.

Emerging trends in modeling techniques

Certain computing and modeling paradigms will see wider acceptance as a result of the widespread use of machine learning models in cutting-edge finance sector applications. Below are examples of a few of these techniques.

Sparsity-aware learning

Many common issues in machine learning have led to the development of a new method of model regularization called sparsity-aware learning. An extensive amount of work has gone into the iterative construction of such frameworks to solve model parameter estimate issues while avoiding overfit. Using iterative strategies like frameworks to estimate model parameters while avoiding overfit is a common approach. Financial modeling applications benefit greatly from sparsity-aware learning algorithms, which provide exceptionally reliable and precise models for a wide range of financial uses.

Reproducing Kernel Hilbert Spaces

A continuous function in the linear space may be evaluated using a Reproducing Kernel Hilbert Space (RKHS). All RKHS functional representations are the limitation of the empirical element showcasing the involved risks, and they are structured as lineal combination of different points of data in the training sets modified by the kernel function, hence they find useful applications in statistical learning. Thus, RKHS has enormous promise in financial risk modeling and appraisal.

Monte Carlo simulation

This modelling algorithm provides the modeller different possible results and the probability that they will occur for any given course of action. It finds use in many fields, including business, energy, project management and monitoring, science and technology, and insurance. To conduct risk analysis, it creates hypothetical scenarios by substituting a probability distribution for a single unknown variable. This method's popularity in contemporary financial modeling is largely attributable to its efficacy in dealing with uncertainty.

Graph theory

Complex problems arise when trying to process and visualize multivariate financial data, on top of the challenges they already provide when modeling. Multivariate financial data may be handled in a way that is both elegant and efficient with the help of graph theory. High-precision modeling of nonlinear and non-Gaussian systems is possible using particle filtering. It is one of the most powerful and widely used modeling tools across a wide range of disciplines, including finance, due to its capacity to accommodate multi-modal input. Particle filtering, in its simplest form, is an approach for identifying the population distribution with minor variance. This is achieved by first determining a collection of randomized samples, which cover different states, then identifying the probability density function, which fits the initial distribution, and lastly replacing the integral operations on the functions with the sample average.

Parameter learning and convex paths

Regularization of optimization techniques has been of paramount significance for the appropriate training of large-scale deep neural networks with millions of parameters, since these approaches have been shown to be particularly successful in training such networks. That's why a lot of effort has been put into calculating how accurate the algorithms are at determining the optimal value of the goal function. When it comes to highly important applications like financial modeling, an estimate of such biases gives the modeler a sense of the degree of error in the models.

Deep learning and reinforcement learning

Most machine learning applications in finance have taken the shape of models based on deep neural network architecture, together with improved methodologies for learning and optimization. In recent years, however, models based on reinforcement learning have made this kind of automation possible. Algo trading, financial asset pricing, share price prediction, and portfolio management are just a few of the many applications that may benefit greatly from the use of reinforcement learning and deep learning frameworks in their development and implementation.

III. AI APPLICATIONS FOR MATERIALS SCIENCE AND ENGINEERING

The study of materials, or materials science, spans a wide variety of academic disciplines. Engineers specializing in materials work to create and improve materials, as well as to discover new applications for them in many other sectors. It was during the Age of Enlightenment that scientists first used the analytical tools of chemistry, physics, and engineering to the study of materials in an effort to make sense of old phenomenological discoveries in the fields of metallurgy and mineralogy. The study of materials continues to combine the disciplines of physics, chemistry, and engineering. Thus, the area was traditionally seen as a sub-field of these linked subjects by academic institutions. It was not until the 1940s that materials science was officially acknowledged as a separate branch of science and engineering, and that's when the world's top technical colleges started establishing departments devoted only to it.

Understanding the role that a material's processing history has in shaping its structure and, by extension, its qualities and performance is a fundamental focus for materials scientists. The materials paradigm describes the concept of processing-structure-properties links. Several scientific disciplines, such as nanotechnology, biomaterials, and metallurgy, all employ this paradigm to further their knowledge. Forensic engineering and fault identification both depend majorly on materials science as they assess the root cause of injuries and accidents, which originate from multifunctioning or defensive goods, components, or buildings. These injuries are fundamental for detecting the causal agent of a wide variety of aircraft occurrences and mishaps.

New materials may be synthesized, and the ability to anticipate different types of chemical synthesis has accelerated the development of ML research in the materials area in recent years. This section will discuss how ML may be used to overcome challenges in the areas of material design, synthesis, and processing.

Accelerated Simulation

Research methods in materials science and computational biochemistry have been stimulated by their third generation. The "structure-performance" computation is the first generation, and it primarily uses the local optimization technique to forecast material performance based on structure. Crystal structure prediction, which predicts structure and performance based on element composition, comes in at number two. The third generation often referred to as "statistically driven design," makes use of ML approaches to forecast the properties of materials based on their physics and chemistry. Unfortunately, the theory's flaws have made it harder to find high-performance materials, and the model's parameters aren't always compatible with real-world scenarios like mixed phase or grain structure. The conductivity of zirconium-doped lithium tantalum silicate, for instance, was predicted to be 10^{-3} S cm^{-1} by DFT, but further measurements have showed that it is more like 10^{-5} S cm^{-1}. Consequently, it is crucial to discover how to apply ML to compensate for the drawbacks of simulation.

Atom2vec

The elements in the periodic table were rebuilt in only a few hours using the unsupervised ML software Atom2Vec. For starters, Atom2Vec learns to differentiate between atoms by examining the substances found in the internet database. Then, we adopt a basic idea from NLP: just as a word's qualities may be inferred from its surrounding words, so too are chemical components grouped in accordance with their immediate surroundings. In addition, the vectorized atom descriptor has a wealth of information on the periodic elementary table, making it a promising new approach to the quantitative description of material data in the years to come.

Increasing Simulation Scale

The appropriate energy or force field may be quickly estimated from theory after ML finds these patterns of recurrence. This is due to the fact that the calculation of the atomic force field exhibits a number of recurring patterns. Extending the timeframe and length of simulation computation may improve the findings from different atoms moving in picoseconds to millions of these atoms moving in nanoseconds. Complex chemical processes and material structures including corrosion, interfacial interactions, and polycrystalline amorphous materials may be simulated using this method.

It is a significant problem to create reliable interatomic potentials for use in large-scaled MDs, interfacial chemical, and surface simulation processes, as a result of the atomic environments and diversified bonds formats. Over the past few decades, interatomic potentials, using ANNs, has been developed for generating the potential energy system’s surface, which is otherwise challenging to define using ordinary potential. In order to verify interatomic possibilities of ANN, Krishnamurthi, Raju, Khambhaty, and Manoharan [5] designed a CuZnO ternary combination model with oxide-based copper cluster as a reference system.

In sum, the possible energy integrated by NNs is significantly precise, with the results being within a small margin of error of the value calculated using the foundational electronic model, and several magnitude orders. The processing requirements for calculating NN potential energies are higher than those of other methods because of the large number of training points needed. At the scales where NN might be beneficial, traditional electronic structure computation faces major hurdles.

Reducing the Amount of Computation

There are so many different permutations of materials that it would take an unreasonable amount of time to do a typical simulation computation to uncover them all. For instance, the smallest known sulfide nanocluster has a bimetallic configuration of Au (SR) that surpasses 32,000, making it a formidable computational task to explore all possible configurations. However, the complexity of computations may be considerably decreased and the filtering velocity could be increased by various magnitude orders if just a subset of data is employed in the training of the ML model and then this model is employed in forecasting these combinations.

To foretell CO adsorption energy of nanoclusters, Magnussen and Burgess [6] recommended a ML framework using the stochastic forest technique. To begin, a training model was constructed using data from DFT simulations of Ag-alloyed

Au nanoclusters. Accuracy rates of 0.78 (R2) and 0.17 (Pearson's r) were achieved in a prediction of the adsorption energy using a two-step procedure for feature engineering and feature selection approach. Hulva et al. [7] observed that the most influential factor in CO adsorption sites was the Ag atoms distribution in Au after deciphering the most significant nodes of random forest. The ML model is readily generalizable to various Au-based nanoclusters. We anticipate the model's primary use will be as a filter to identify promising sources for in-depth study.

Predicting the Property of New Materials

Jiang et al. [8] aim to optimize properties such as electrolyte conductivity, thermodynamic Seebeck coefficients, and the efficiency of power conversion of organic-inorganic perovskite. Even after extensive trial and error, results from numerical simulations or the perception of chemical experts are sometimes disappointing. Applications of ML models, which can forecast the characteristics and structures of materials with a reasonable degree of accuracy prior to synthesis, may therefore be of great use. The ML model built in MATLAB was utilized by Benea and Benea [19] to locate rare solid electrolytes amongst over 12,000 different substances. Next, they looked through the research to find 40 crystalline solids with known atomic structures that matched a popular set of electrolytes utilized in exercise. Despite the relatively small quantity of the dataset, the "smart" characteristic based on past physical knowledge is required for an accurate representation of the data.

In order to characterize the local organization of atoms and crystalline chemical features, the author uses the atomic structure downloaded from ICSD as inputs and computes approximately 20 features according to their atomic geography, electro-negativity, mass and atomic radii of the structural element, such as lithium bond iconicity, atomic volume, lithium neighbouring components, and the minimum anion-to-anion separation extent.

The values of the experiment with the conductivity of lithium ions are applied as outputs, while a list of 40 known materials serves as inputs to a machine learning method. By adjusting the model's parameters in a steady fashion, it is possible to screen and categorize solid electrolytes, and eventually 317 compounds were indicated as promising. The updated MATLAB model proved to be two times as effective as Stanford graduate students working in a similar area, and three times as effective as random guessing, when it came to finding prospective new materials. The F1 score is almost 50% lower than the DFT values..

High-throughput simulations are just as useful as practical testing for ML training data. The thermodynamic stability of double perovskite halides was investigated by Dimopoulos et al. [10] utilizing high-throughput computation and ML. They began by creating a high-throughput DFT-based breakdown energy database, which was then linked to the thermodynamic constancy of 352 potential perovskite materials. They used this data to teach a machine learning model. Experiments on perovskite formability of 246 A_2B (I) B (III) X_6 elements (F1 95% score) further corroborated its prediction performances. The results of this study demonstrate that the model of prediction in ML is both cost-effective and efficient compared to experimental approaches.

Many important materials, including organic-inorganic lead-free hybrid perovskite, LEDs, OLEDs, LEDs, and monoatomic catalysts have been designed using similar approaches. The experimental validity of the latter two approaches has also been shown. In its current form, material science still relies somewhat on tried-and-true methods. There are still some hypotheses being employed to cut down on the amount of needed tests, and this trend is only expected to increase in importance. Instead, the regression model may be used to narrow down a pool of potential materials to only one that shows promise, cutting down on wasted time and effort spent on trial-and-error procedures.

Synthetic Route Planning

As organic synthesis follows a tried-and-true protocol, Haggin [11] may easily develop software to solve complex synthesis-related challenges. In the eyes of computer experts, chemical reactions are just a bunch of numbers that represent the link between two substances. Data structures like graphs and networks may be used to represent this presence. This structural information might then be used by AI to direct the synthesis process. Using a feedback loop and realtime spectrum analysis, Yang, Hansen, and Baldi [12] showed a robot that can carry out six separate organic synthesis experiments concurrently.

A chemically-assembled pressure pump and a raw-materials storage tank are its core components. The six reaction bottles that run in parallel are fed reactants by these pumps. In addition, the robot employs NMR and IR spectroscopy for real-time evaluation of reactions and SVM algorithm to routinely classify the reactions mixtures into both non-reactive and reactive combination. This approach can anticipate the reaction of reagent combinations and is quicker than manual experimentation. Moreover, the robot found four additional responses and had a prediction accuracy of over 80% after only gathering data from about 10% of the experimental dataset.

Experimental Parameter Optimization

In conventional material development, many factors in synthesis, processing, and device assembly must be examined and changed by hand. The efficiency is quite poor, and it may not even be possible to determine the best settings. To pinpoint the optimal value in the vast parameter space, ML makes use of its potent nonlinear regression capabilities. The concept has been implemented in welding procedures. Recently developed friction stir welding (FSW) has found widespread use in sectors as diverse as aerospace, shipbuilding, automotive manufacturing, and more. The impacts of the traditional parameters of welding, such as the maximum shear strain on the tool spin, strain rates, torque, temperature, and possible causal agents on void formations, were investigated by Habibizadeh, Honarpisheh, and Golabi [13], who collected about 107 independent data for experiments from a wide-range authoritative literatures for the purpose of training different model of ML, including decision trees, and the neural networks. The findings demonstrate that both algorithms are capable of making reliable predictions about the emergence of errors, with a maximum prediction precision of 96%. The model permits for a complete optimization of the parameters of the welding process, preventing adverse phenomena like void development in FSW from Occurring.

Examples like this have been used in 3D printing. Microelectronic devices on flexible substrates are often fabricated using AJP (aerosol jet printing), a non-contact three-dimensional printing method. While it has the capacity to deposit unique designs, the printing quality will be heavily influenced by the intricate interplay between the primary process factors. To find the optimal AJP process window across various design domains, Caldana et al. [14] suggested a novel hybrid ML approach. Data clustering, knowledge transfer, classification, and experimental sampling are all staples of traditional ML approaches. The technique employs the experimental design known as Latin hypercube sampling to exhaustively probe the two-dimensional design spacing at a specific printing velocity. Then, using the K-means clustering technique, we studied how the SHGFR and CGFR affected the printing line quality, and we used a support vector machine to zero in on the sweet spot for running the presses (see **Fig 4**). The transfer learning technique exploits the association between several operation process windows in order to accurately detect more of these windows at varying printing rates. As a result, the improved printing speed allows for a much smaller sample size of rows to determine the new operating process window.

Lastly, an incremental categorization approach is employed to generate a 3D operating process windows that strike a balance between CGFR, printing speed, and SHGFR. Instead than relying on trial-and-error experimentation, as is the case with traditional approaches to quality improvement in 3D printing, this new approach is grounded on the principles of data mining and knowledge discovery. As a result, all of the information needed to maximize printing line quality may be gleaned from the many design spaces that exist.

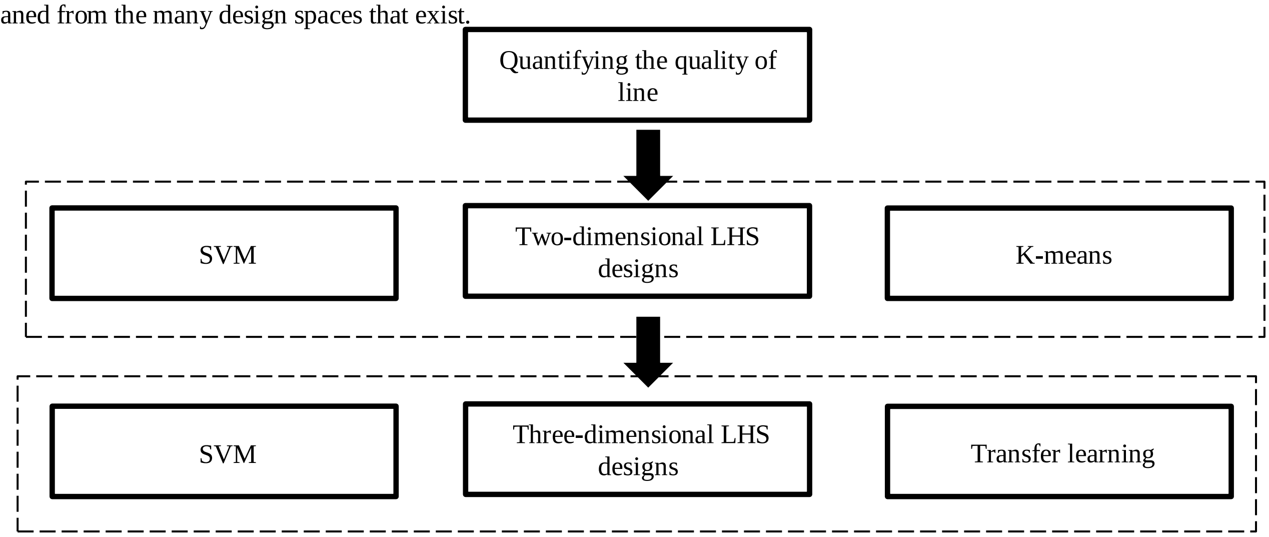

Quantifying the quality of line
SVM
Two-dimensional LHS designs
K-means
SVM
Three-dimensional LHS designs
Transfer learning

Fig 4. High-level Overview of the Hybrid ML Approach to Printing Parameter Optimization

As material synthesis moves toward complete automation, it will be included into industrial production 4.0 alongside tools like polymers' computable high-throughput analysis platforms. To begin this kind of high-throughput analysis successfully, ML must first analyse parameter spaces to identify the raw material's optimal ratio and the catalyst delivery rate for synthesizing organic molecules with the desired molecular weight, distribution, and side reaction profile.

Upgrading of Characterization Methods

Due to improvements in representational technologies, scientists have been able to examine structures and motions on the atomic scale, leading to the discovery of new rules governing materials. High-throughput analysis and preparation of materials based on the application of AI will be a given once the Material Genome Project matures. Due to deep learning's effective use of convolutional neural networks, significant development has been witnessed in the field of picture recognition.

This skill and knowledge in pattern recognition may be simply applied to the characterisation of micromaterials in images. Material science relies heavily on electron microscopy and defect analysis because they provide invaluable information about the microscopic structures and behaviour a wide range of materials systems and bio-materials. Establishing an adaptable and robust framework for automatic fault detection and categorization in electron microscopy would allow for faster analysis both post-recording and during picture capture. Manual identification is still the norm, despite being time-consuming and prone to error. This is because a higher number of images is needed to obtain statistical insights. As of late, Chakroun, Bouhamed, Kallel, Solaiman, and Derbel [15] used ML, image analysis, and computer vision methods to learn about the magnitude and nature of faults (see **Fig 5**). The current performance of the application agrees with the quality analysis performed manually. If the software is upgraded, it will soon be possible to analyze massive data volumes in real time.

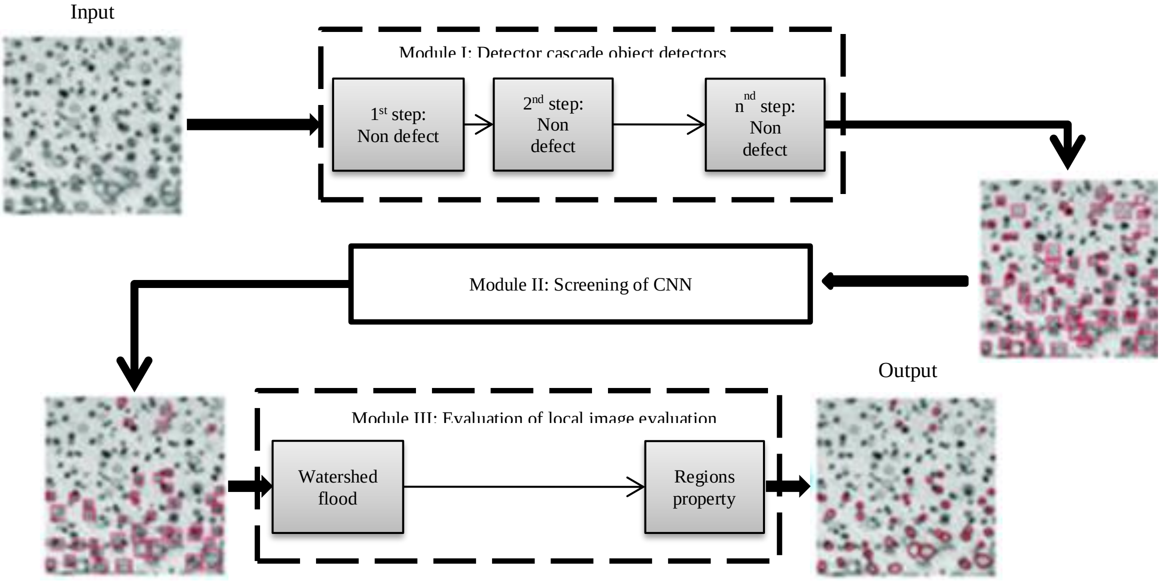

Input
Module I: Detector cascade object detectors
1st step: Non defect
2nd step: Non defect
nnd step: Non defect
Module II: Screening of CNN
Output
Module III: Evaluation of local image evaluation
Watershed flood
Regions property

Fig 5. Proposed Automated Detection Method

Fig 5 is a high-level diagram depicting the suggested automated detection method. The first step in processing an incoming micrographic picture is the Cascade Object Detector, followed by the Convolutional Neural Network (CNN) detection and lastly the LIA (Local Image Analysis) module. When module I is completed, module II is then used to refine the bounding boxes and loop location, hence reducing false positives. Artificial intelligence (AI) may also be used to evaluate data from X-ray diffraction (XRD). When dealing with massive amounts of high-throughput characterization data, performing individual analyses to extract representative samples is a laborious and time-consuming process. Scientists may use ML to increase the speed of their analyses and unearth previously unknown patterns in their data.

Min [16] deposited ternary FeGaPd compound films on a single silicon wafer, resulting in 535 samples with varying ternary FeGaPd compositions, each measuring 1.75 mm^2 in area. X-ray diffraction (XRD) characterisation data was collected for 273 samples. Then, ML is used to perform unsupervised clustering on 273 XRD sample data using a hierarchical clustering technique, with the goal of combining as many single-phase samples into a single cluster as feasible. The analysis is far more efficient since just a small subset of the data in each cluster is examined. Based on these findings, it is clear that ML's dimensionality reduction and clustering technique may be used to effectively evaluate high-throughput XRD dataset, identify the stage distribution pattern and the multi-phase junction, and speed up the discovery process.

Increasing the number of times you cycle a lithium-ion battery will reduce its capacity. Researchers into battery technology have always been interested in improving cycle life. A new, data-driven, large-scale model has been constructed by Sandie et al. [17]. The capacity to employ ANNs to assess the laws of high-dimensional dataset allows for the prediction of the whole lifecycle of graphite batteries, or commercial lithium iron phosphate using just the charging and discharging dataset of the first couple of cycles, without any analysis of the process of battery deterioration. The first 100 cycles are used in the regression setup, and the prediction error is reduced to 9.1%. Using data from the first five cycles, the author constructs a classification scheme that yields a prediction error of just 4.9%. As a result, there will be fresh chances to improve battery manufacturing, cascade use, and optimization. Manufacturers of rechargeable batteries may, for instance, shorten the time it takes to create new battery types, verify the viability of proposed production methods, and categorize batteries in terms of their expected lifespan. In a similar vein, the expected lifespan of batteries in consumer electronics may be estimated. As a whole, the work places emphasis on the integration of data collection and data-driven modeling, an approach with promising implications for the study and improvement of complex systems like lithium-ion batteries.

The ambiguity associated with analyzing impedance data may also be cleared out with the use of ML. When it comes to studying and diagnosing electro-chemical battery and possible electro-chemical storage systems for energy distribution in the future, electro-chemical impedance spectroscopy (EIS) is a significantly effective technique. On the other hand, analyzing a huge quantity of EIS data is notoriously challenging. Many optimization techniques have gaps in their functionality. To achieve the right convergence of the fitting, Researchers must carefully create the equivalent circuit (EC) model, choose acceptable beginning values for the parameters of each component, and continually validate the output during the process. For this reason, Altay and Yildiz [18] recommended a ML-inverse approach that condensed a million separate fitting optimization issues into a solitary one. Using many perspectives from the ML literature resulted in an error rate of less than 1% for addressing a single optimization task. The physical model parameter could be effectively fitted to the gathered set of data if an open-source model is constructed for FIS testing, and the system can be simply adaptable to other impedance spectra. There is no need for human intervention and the results are consistent every time. You may get a copy of the code used in this study by visiting -https://github.com/samuel-buteau/eisfitting. This self-deprecation of material science research as "stir-fried meals" is a current problem. It uses trial and error to find new materials, like as salt and water, to add to existing ones. High-throughput computers and machine learning allow material scientists to reduce the time and effort spent on trial and error.

Free open-source software that integrates AI data processing capabilities with a user-friendly interface may be necessary for the development of tangible AI in the future. Artificial intelligence has the potential to monitor all areas of scientific inquiry and provide other analyses to address representational issues. For the benefit of other researchers trying to address the same problems, researchers may post their own experimental procedure and findings. The use of AI will not eliminate the need for synthetic chemists, though. Indeed, AI will definitely become much significant to chemists to enable them to determine synthetic pathways better and promptly, but synthetic chemists will continue to discover novel reactions through actual scientific evaluation and diversify the theoretical framework of Chemistry. Using ML technology in conjunction with existing experimental data and theoretical foundations, AI-assisted materials design, application, characteristics, and synthesis research will significantly enhance the research effectiveness of scientists in the materials field and aid in rapid advancement of materials science.

IV. CONCLUSION AND FUTURE RESEARCH

This article provides a summary of the current state of the art in materials AI research, focusing on its most salient applications and benefits over more traditional approaches. High-throughput experimentation, high-throughput characterization, and high-throughput simulation computations are all necessary for the further growth of material informatics. In this section, we shall discuss the software and hardware future. Data analysis (statistical approach) is at the heart of ML, and the necessary data strives for breadth, depth, and impartiality. The calculated parameters of previous material informatics investigations lacked sufficient precision, limiting the scope of the research. The difference will be substantial if the datasets are made of more precise experimental findings. However, due to the overwhelming concentration of attention-grabbing research hubs, the existing experimental samples are insufficient. Generative adversarial networks, active learning, transfer learning, and auto-encoders are just a few examples of models that function well with little data. To prevent the "Black Box" problem, ML frameworks have to be changed into physical visuals or real-time knowledge. It may be possible to get some insight by averaging the responses of the neurons to the various descriptions. Materials informatics might be advanced by the use of other explanatory frameworks (for example decision trees) that can show the importance of different elements through the relative weight of their branches and nodes.

In most cases, a large amount of data is required for successful training of ML models. There are several potential sources for this kind of information, including web databases, published articles, and high-throughput experimental equipment. The use of online databases like ImageNet is becoming more common in the implementation of deep learning. Similar infrastructure is required for the advancement of material informatics. To compile data on electrolytes such their ionic conductivity, transference number, and chemical stability, Researchers created a database. In addition to a wealth of information on materials, published papers are a great source. Once these publications are organized by defined article formats, Researchers may conduct targeted information searches with ease using natural-language-processing technologies. High-throughput synthesis and characterisation tools may be upgraded by adding more sensors and software. The data acquired by these instruments is sent back into AI models for use in fine-tuning trial conditions. The settings may then be tweaked to generate samples with the desired characteristics. By these efforts, the "composition-structure-property-processing-application" link will be mapped in materials informatics.

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]. W. Sha et al., “Artificial intelligence to power the future of materials science and engineering,” Adv. Intell. Syst., vol. 2, no. 4, p. 1900143, 2020.

[2]. B. G. Hyde and S. Andersson, Inorganic Crystal Structure. Nashville, TN: John Wiley & Sons, 1989.

[3]. M. Pentz, Handling Experimental Data. Buckingham, England: Open University Press, 1988.

[4]. J. X.-Y. Lim, D. Leow, Q.-C. Pham, and C.-H. Tan, “Development of a robotic system for automatic organic chemistry synthesis,” IEEE Trans. Autom. Sci. Eng., vol. 18, no. 4, pp. 2185–2190, 2021.

[5]. P. Krishnamurthi, Y. Raju, Y. Khambhaty, and P. T. Manoharan, “Zinc Oxide-Supported Copper Clusters with High Biocidal Efficacy for Escherichia coli and Bacillus cereus,” ACS Omega, vol. 2, no. 6, pp. 2524–2535, 2017.

[6]. S. Magnussen and D. Burgess, “Stochastic resampling techniques for quantifying error propagations in forest field experiments,” Can. J. For. Res., vol. 27, no. 5, pp. 630–637, 1997.

[7]. J. Hulva et al., “Unraveling CO adsorption on model single-atom catalysts,” Science, vol. 371, no. 6527, pp. 375–379, 2021.

[8]. Z. Jiang et al., “Ag/Br dual-doped Li6PS5Br electrolyte with superior conductivity for all-solid-state batteries,” Scr. Mater., vol. 227, no. 115303, p. 115303, 2023.

[9]. M. L. Benea and O. D. Benea, “Mathematical modelling in Matlab of the experimental results shows the electrochemical potential difference - temperature of the WC coatings immersed in a NaCl solution,” IOP Conf. Ser. Mater. Sci. Eng., vol. 106, p. 012025, 2016.

[10]. E. A. Dimopoulos et al., “HAYSTAC: A Bayesian framework for robust and rapid species identification in high-throughput sequencing data,” PLoS Comput. Biol., vol. 18, no. 9, p. e1010493, 2022.

[11]. J. Haggin, “Easier-To-Use Computer Systems Available to Organic Chemists: New SYNLIB and Chemist’s Personal Software Series systems are usable with personal computers, aim at synthesis chemists,” Chem. Eng. News Archive, vol. 63, no. 34, pp. 13–16, 1985.

[12]. Y. Yang, L. Hansen, and A. Baldi, “Suppression of simultaneous Fmoc-his(trt)-OH racemization and Nα-DIC-endcapping in solid-phase peptide synthesis through design of experiments and its implication for an amino acid activation strategy in peptide synthesis,” Org. Process Res. Dev., vol. 26, no. 8, pp. 2464–2474, 2022.

[13]. A. Habibizadeh, M. Honarpisheh, and S. Golabi, “Effect of friction stir spot welding parameters on the microstructure and properties of joints between aluminium and copper,” Weld. World, vol. 66, no. 9, pp. 1757–1774, 2022.

[14]. A. C. F. Caldana, J. H. P. P. Eustachio, B. Lespinasse Sampaio, M. L. Gianotto, A. C. Talarico, and A. C. da S. Batalhão, “A hybrid approach to sustainable development competencies: the role of formal, informal and non-formal learning experiences,” Int. J. Sustainability Higher Educ., vol. 24, no. 2, pp. 235–258, 2023.

[15]. M. Chakroun, S. A. Bouhamed, I. K. Kallel, B. Solaiman, and H. Derbel, “Feature selection based on discriminative power under uncertainty for computer vision applications,” ELCVIA Electron. Lett. Comput. Vis. Image Anal., vol. 21, no. 1, pp. 111–120, 2022.

[16]. H. S. Min, “A review on the Penternary compound thin films,” Researchgate.net. [Online]. Available: https://www.researchgate.net/publication/317637819_A_review_on_the_Penternary_compound_thin_films. [Accessed: 03-Mar-2023].

[17]. A. B. Sandie et al., “Observed versus estimated actual trend of COVID-19 case numbers in Cameroon: A data-driven modelling,” Infect. Dis. Model., vol. 8, no. 1, pp. 228–239, 2023.

[18]. A. Altay and A. B. Yildiz, “Complete electrical equivalent circuit based modeling and analysis of permanent magnet direct current (DC) motors,” WSEAS Trans. Circuits And Syst., vol. 21, pp. 182–187, 2022.