Ceramics Research Done Smartly: Machine Learning for Accelerating the Transition to Sustainable Ceramics

Myungbeen Choi

Received June 15, 2024
Accepted September 23, 2024
Electronic access October 15, 2024

Ceramics have been a product of human ingenuity throughout history. From their usage in pottery and artifacts in the past to industrial, electrical, and aerospace applications of ceramics in the present, ceramics are ubiquitous in our lives. Unfortunately, manufacturing ceramics is extremely energy-intensive and is responsible for more than 8% of global anthropogenic CO_2 emissions. There have been continuous efforts to make these valuable materials more sustainable. Recently, with the advance of data science, organizing and learning from data has become much less time-consuming. This literature review aims to coalesce the two fields, sustainability and data science, with a focus on ceramics. Instead of narrowing our focus to specific ceramics, such as glass or cement, we proposed procedures that can be broadly applied. By analyzing different data science approaches and their applicability to creating ceramics databases, we evaluated various ways to promote ceramics sustainability using data science. More specifically, we analyzed vital insights or techniques that may be applied elsewhere and sought opportunities to create procedures that required the least human input and benefited the most from machine learning models. We categorized different approaches toward regression and data extraction, examined suitable approaches for various scenarios, and qualified these efforts by explaining the impact new databases can have on the sustainability of ceramics. We hope this literature review will prove relevant to readers with an interest in data science, sustainable ceramics, and their intersections.

Introduction

Ceramics are defined as materials that are neither metallic nor organic. These materials often present themselves with valuable properties, such as heat resistance, semiconducting behavior, or inertness. For example, refractories are used in extreme-temperature industrial processes because of their high resistance to heat, and technical ceramics are used in chemical processing plants because of their chemical inertness [1]. Ceramics are undoubtedly valuable, but unfortunately, producing ceramics is a highly energy-intensive process. Aside from the mechanical energy needed to mill and pulverize the mined raw materials (e.g., rocks of clays, silicas, limestones), these raw materials require incredibly high temperatures (e.g., above 1500 °C in specially equipped furnaces) to react in solid or molten form and yield the solid solutions that define the products, and these processes are often carried out using natural gases, which account to around 82-95% of the total energy usage [2]. The energy requirements of a simplified list of common ceramic products are listed below in Table 1.

Type of Product	Energy Requirements (per ton of product)
Clay and Ceramic Floor Tiles	940 kWh
Ceramics for Electrical Use	5000 $\sim$ 5830 kWh
Bricks	380 kWh
Roof Tiles	1250 kWh

Table 1: A Simplified Subset of Common Ceramic Products and Energy Requirements [2]

These energy requirements quickly add up to a significant amount of natural gas and electricity consumption. For example, in 2007, the ceramics manufacturing sector was responsible for 12.3% of the total natural gas consumption in Turkey [3], and in 2014, the ceramic industry was accountable for 5.8% of the industrial energy consumption in Brazil [4]. Not only do ceramics require an abundance of energy, but they also produce tons of greenhouse gas emissions. The associated greenhouse gas emissions are not attributable only to the energy input to the manufacturing process (e.g., process heat, electricity, mechanical power) but also to the process emissions emanating from the solid-state reactions responsible for chemically converting the raw materials to products, accentuating the climate impact of the ceramics industry via these process emissions. Still, the majority (more than 80% [2]) of the greenhouse gas emissions are attributed to the firing and drying stage, generating around 265 kg CO_2 per tons of fired ceramic tiles [5], the product with the highest production value in the European ceramic industry [1]. These emissions from the European ceramic industry amount to 19 million tons of CO_2 annually [1], and much research is necessary to make the ceramics industry more sustainable.

We provide a general schematic of the ceramics industry below to better explain the sustainability challenge in ceramics. After mining from Earth deposits, raw materials such as clay, silica, or limestone are transported to and stored in the manufacturing plant. This transportation process could lead to significant emissions if raw materials are mined far from the ceramic plant. The raw materials at the ceramic plant are then milled, mixed, and pulverized to increase the surface area available for the ensuing solid-state reactions. Often, they get formed into a paste and shaped into a mold before getting dried, to prepare them for the firing step. The raw materials then react under firing in furnaces, where heat management and distribution are of paramount importance to the quality of the product. Afterward, finishing often includes cooling, packaging, and regular testing for quality control of the products.

Figure 1 overlays some examples of opportunities for sustainability in existing ceramic plants. Following the schematic below, one can start by recycling ceramic products to prevent the extraction of unnecessary raw materials. In fact, the European Ceramic Industry Association envisions a circular pathway of ceramic materials in which refuse from manufacturing and distribution sites or broken pieces of ceramics from the "grinding and the decorating and glazing operations"[2] are used as fillers to minimize new raw materials in the production mix, solving both a waste problem and an emissions problem. Some wall and floor tiles industry manufacturers transitioned from rotary printing to digital printing and created walls and floor tiles with 80% recycled ceramics using ceramic inks instead of decorative pastes. Clay pipes consisting of 40% new raw materials and refractories containing 20% to 80% new raw materials can be produced, and it is possible to reuse 90% of expanded clay products[1]. Currently, "around 30% of the materials used in the ceramic industry are dumped in landfills"[2]. To decrease that number further and reuse more ceramics, there must be technological breakthroughs that allows the production of ceramics using higher proportions of recycled pieces while retaining the efficiency.

Continuing the schematic, one can minimize transportation energy costs and emissions by placing the ceramics manufacturing site near the mining and raw materials extraction site. Although the initial effort to make this change does require energy, if the demands for ceramics persist, the energy saved and emissions reduced from not having to travel long distances carrying heavy raw materials will outweigh the initial energy cost and emissions in the long run. A potential barrier that can prevent this change is the "recurrent lack of information" provided to the manufacturers[2]. The absence of a detailed "cost-benefit and viability analyses"[2] of these energy-efficient practices can generate uncertainties for investment, which will prevent manufacturers from implementing these changes.

Next on the schematic is the drying step, which is crucial as this step controllably gets rid of water and volatile organic impurities contaminating the raw materials, ensuring the mechanical integrity of the fragile and, as of yet, wet shapes. Here, there exists an opportunity for heat integration, where the flue gases from the firing step heat the drying air or are applied directly to dry the incoming feed to the furnace. Using environmentally sustainable fuel sources such as green hydrogen during this firing stage could significantly improve the sustainability of the overall process. The furnace can also be lined with ceramic refractories that more effectively trap the heat inside, offering improved insulation. Further, capturing the carbon released during the drying and firing stages can help decarbonize the production process. D. Rio and coworkers propose some emerging future technologies that can help decarbonize this manufacturing step of ceramics, which include vacuum drying, microwave-assisted drying and firing, a hybrid kiln that allows the selection of either electric heating or primary fuel, and fast-firing cycles that reduce the firing temperature by up to 50°C[2]. A limitation to these changes proposed for the drying step is that the potential economic disincentives[2] arising from these changes or the switch to using electricity and green energy sources can be a burden for manufacturers that must be alleviated over several years. For example, it is unreasonable to expect carbon capture and storage technologies to be implemented in competitive industrial settings, considering the impractical cost of them. Not to mention, the current technology is far from sufficient to implement some of these changes. For example, electric kilns "have not yet been implemented on a continuous and large scale", which implies that using electricity during the drying and firing stage isn't possible as of now[2].

Some additional measures the European Ceramic Industry Association proposes include "the reduction of carbon-containing additives", "optimization of the carbon content of clay mixes", and "offsetting measures"[1]. Carbon-containing additives can often improve the properties of ceramics, allowing them to be used for new and practical purposes. For example, adding carbon additives to ZrB_2 ceramics can improve the "oxidation resistance, mechanical and fracture–toughness properties"[6]. To incentivize reducing carbon-containing additives, new replacements and substitutes with less emissions must be developed through sufficient research and testing, and optimizing the carbon content of clay mixes also requires further research. Considering the apparent limitations and barriers towards decarbonization in conjunction with the fact that the ceramics industry remains intrinsically emissions-intensive and geared towards end products, it is pivotal to offset these emissions by producing advanced ceramics that are energy and resource efficient and can be used for green technologies, to reach the vision of a net-zero emission rate proposed by the European Ceramic Industry Association[1].

Despite the propitious options to make ceramics sustainable,

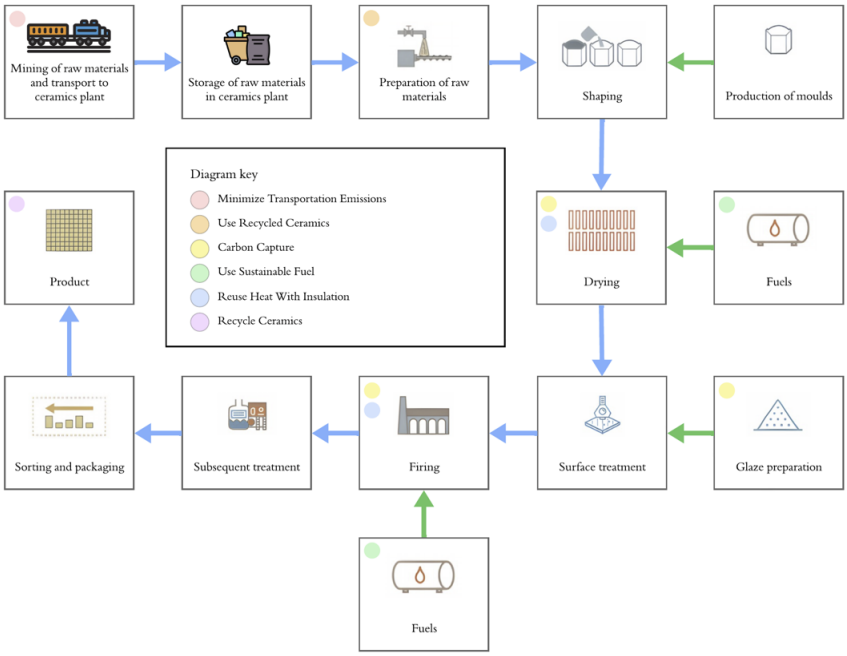

Figure 1: A Simplified General Schematic of the Ceramics Manufacturing Process adapted from Ceramic Roadmap to 2050

there are clear sociotechnical limitations that hinder the complete decarbonization of ceramics. Particularly, further research is necessary to pursue technological measures such as producing advanced ceramics or replacing carbon-containing additives. However, by leveraging data-driven strategies, we can begin to address these challenges more effectively. Here, we suggest that a smart approach towards sustainable ceramics research is data-centered, including building, maintaining, and seamlessly interacting with comprehensive databases that adhere to FAIR data principles (findable, accessible, interoperable, and reusable).

To make this point, we highlight two recent articles discussing the highly successful results of this endeavor. The first is by K. Gong and coworkers at Princeton University, where they established a strong research program geared towards the manufacture of cement with less emissions, by designing alkali-activated kaolin clays as raw materials that help turn industrial wastes and already-calcined clays (which emit no further CO_2 upon processing) into cementitious binders, minimizing process emissions[7]. Specifically, they applied computationally intensive density functional theory simulations to calculate the binding energies of seventeen such alkali cations to aluminosilicate dimers and trimers and then correlated these binding energies to two other properties that are readily obtainable (i.e., do not require DFT simulations): the ionic potential and field strength of these cations. The high-quality second-order polynomial correlations (R^2=0.99-1.00) highlight their usability to rapidly estimate such binding energies of new cations or new substrates, extending this research's impact into materials not initially simulated by the authors[7]. The data on binding energies of different cations to different clay materials can help design alkali-activated kaolin clays as raw materials by providing insight into the performances (e.g., performance against sulfuric acid attacks) of the resulting sustainable cement. This endeavor highlights the importance of a data-centered approach toward innovative ceramics research.

The second article is by S. Kim and coworkers at Seoul National University. They successfully created an extensive band-gap database for semiconducting inorganic materials, as applied in photovoltaic devices. By using a hybrid functional and considering a stable magnetic ordering, they made a database for band gaps for 10,841 materials with a significantly smaller root-mean-square error of 0.36 eV than the 0.75-1.05 eV of existing databases[8]. Additionally, they identified a considerable number of small-gap materials that were misclassified as metals in other databases. By correctly classifying these materials as semiconducting inorganics, they provided a broader selection of materials that manufacturers or researchers may consider using. This can come with various benefits, such as reduced costs of materials, reduced transport emissions, and more. Overall, S. Kim and coworkers created a highly impactful (and highly cited) database utilizing a data-centered approach, further highlighting

the potential of a data-centered approach for scientific research and innovation. Band gaps are but one of a myriad of properties that define the utility of ceramic materials in advanced and sustainable applications (see Table 2 for a non-comprehensive list of such properties and applications).

Lamentably, there is a severe lack of comprehensive, reliable, up-to-date, and accessible ceramics property databases[9], a problem that must be addressed to transition to sustainable ceramics efficiently. This paper suggests combinations of already-present machine learning methods to maximize the accuracy of ceramics property databases that can be extracted with minimal human input, current literature, and available data.

Results

An Overview of Machine Learning Models

Following the recent rise of machine learning, there have been recent efforts to extract data from text using natural language processing models. For example, M. Polak and coworkers proposed a successful procedure using ChatGPT to extract mid-sized databases from text[10]. This process requires minimal prior knowledge of coding and the property being analyzed, but can generate databases with high precision and recall in a single day. This automatic quality of machine learning, being able to extract data from literature or learn from available data to create sufficiently accurate databases with minimal human input, is what makes machine learning so valuable and, thus, will be the focus of this literature review.

In the world of data science, machine learning methods have varying types of training data and tasks to solve. Supervised models or algorithms refer to models trained on data sets with labels -— either categorical or continuous — corresponding to the data points, and can help predict material properties. Some supervised models, such as graph neural networks, can input graphs as data, which is useful when predicting properties based on molecular structures that are represented as graphs. In contrast, unsupervised models are trained on data sets that lack such labels and are often used for clustering or anomaly detections, but lack utility for the purpose of creating databases for ceramics data. This paper will focus on the following three most common types of supervised machine learning models that were present in the literature surveyed: Support Vector Machines, Decision Trees, and Neural Networks. A brief overview of these three types of models is described in Figure 2.

Support Vector Machines

Support Vector Machines (SVMs) are supervised learning algorithms that can be applied for both classification and regression. For classification tasks, the machine finds the most optimal decision boundary, referred to as a hyperplane, that maximizes the margin, or distance, between the nearest data points of two different classes and the decision boundary. A larger margin implies a clearer separation between the two classes of data and, therefore, a better generalized classifying model. An SVM for regression, often referred to as Support Vector Regression (SVR), is an extension of SVM that predicts continuous variables. Instead of finding a decision boundary that maximizes the margin, an SVR finds a function that predicts the output value within a specified margin of tolerance[11].

Decision Trees

Decision Trees are supervised learning algorithms that utilize tree-like structures to make predictions. These trees are composed of nodes and branches connecting the nodes. The starting node at the top of the tree is called a root node, where the input is given. The nodes at the bottom of the tree are called leaf nodes, and these nodes represent all possible outcomes within the dataset or scenario. Starting from the root node, at each node, the machine makes a decision based on its training of previous data points and sends the input to a new node that was connected to the previous node via a branch. This process is then repeated until the input arrives at a terminal node representing a possible output. There are various types of decision trees, including Random Forest, Gradient Boost, Extreme Gradient Boost, and Adaptive Boost decision trees, all of which are commonly used for regression. Figure 3 below is a brief overview of these four special types of decision trees for regression[12, 13, 14].

A random forest decision tree is an ensemble of decision trees trained separately on random samples of the training data. These independent decision trees are run in parallel when given an input, and the individual results generated by each of these decision trees are aggregated to make a final prediction. A gradient boost decision tree is a series of several decision trees that attempt to predict the errors from the previous decision trees in the series. An extreme gradient boost decision tree is a gradient boost decision tree with regularizations that prevent complex models that are overfitting the training data and/or aren't generalizing well. These regularizations include but are not limited to, lasso regularizations, which encourage sparsity, and ridge regularizations, which reduce the variance of the model. An adaptive boost decision tree consists of a group of independent decision trees that learn adaptively, forcing decision trees that made significant errors to train more on those instances to reduce the large errors. The results from these independent decision trees are then weighted and compiled so that the decision trees that made those significant errors have less influence over the result[12, 13, 14].

Neural Networks

A Neural Network is a program or model that intends to replicate how our brain makes decisions. The nodes in a neural network are connected to one another and serve as neurons of the machine. The individual nodes input data, calculate a weighted output using the input data, and compare the output to a specified threshold value to "make decisions"[15]. When these models are trained on large data samples, they can classify, cluster, and

Table 2: A Simplified Subset of Major Property Categories for Technical and Sustainable Ceramics

Property	Definition	Application	Type	Unit
Thermal Conductivity	Ability to conduct heat	Insulation, electronic cooling	Thermal	W / m · K (Watts per meter-Kelvin)
Band Gap	Energy difference between the highest valence bond and the lowest conduction band	Semiconductors, solar cells, LEDs	Electronic	eV (electron volt)
Piezoelectricity	Ability to generate an electric charge in response to mechanical stress, or vice versa	Sensors, actuators, transducers	Electrical/Mechanical	Vm/N (voltage-meter per Newton)
Refractive Index	Measure of light bending in a material	Lenses, photonic devices, optical networks	Optical	dimensionless
Process Emissions	Emissions generated during manufacturing or industrial processes	Ceramic (cement) industry	Manufacturing	kg CO_2/kg product (kilogram CO_2 per kilogram product)
Corrosion Resistance	Ability to withstand degradation by chemical reactions	Chemical processing	Chemical	mm/yr (millimeter per year)
Melting Point	Temperature at which a material changes from solid to liquid	Material processing, casting, welding	Thermal	°C or K (degrees Celsius or Kelvin)
Fracture Toughness	Ability to resist fracture when a crack is present	Structural ceramics, dental implants	Mechanical	Mpa (megapascal times the square root of meters)
Oxygen Ionic Conductivity	Ability to conduct oxygen ions under an applied voltage gradient	Solid oxide fuel cells, oxygen sensors	Electrical	S/m (siemens per meter)

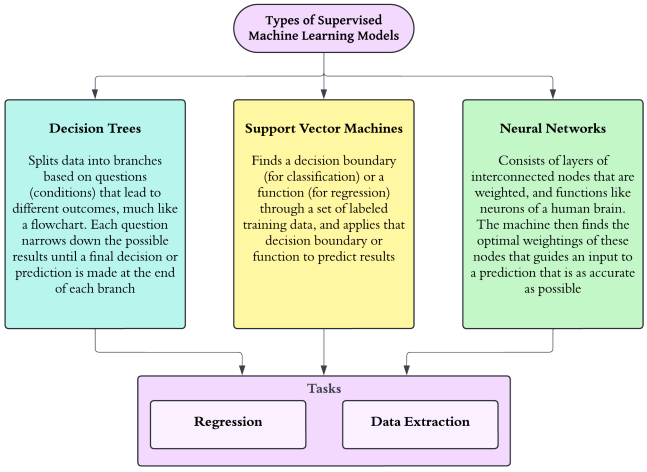

Figure 2: A Simple Overview of Three Types of Machine Learning Models

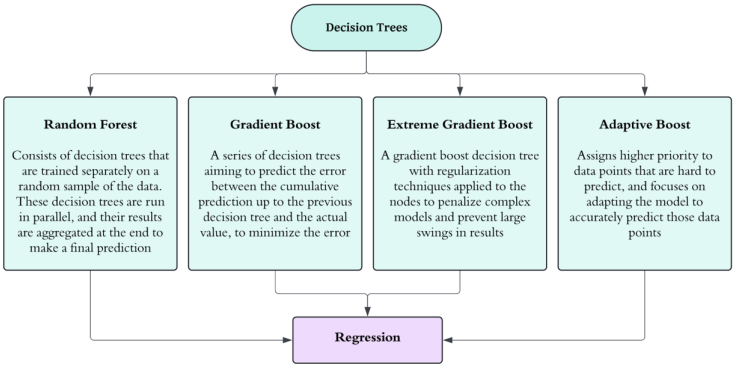

Figure 3: A Non-Comprehensive Diagram Depicting Various Decision Trees for Regression

estimate data rapidly and with high accuracy. Figure 4 is a diagram depicting common types of neural networks and a non-comprehensive type of tasks these models can complete.

Graph Neural Networks are a type of neural network that can intake graph-structured irregular data. These neural networks are especially useful in understanding materials because they can input molecular structures as data and handle them efficiently. Deep Neural Networks, commonly referred to as Deep Learning, are neural networks with three or more layers, which is common for models that require complexity. These models can be trained to estimate material properties using other more available properties without knowing the specific relation between those properties beforehand. A Large Language Model (LLM) is an application of deep learning that utilizes multiple layers of deep neural networks for natural language processing and is trained on extremely large samples of textual data. These various types of neural networks can be useful in ceramics research because of their ability to interpret large amounts of data or text, to predict numerical data, or to extract different data from the vast sea of literature[16].

Machine Learning Models Used for Regression

Machine learning models such as deep neural networks and the various decision trees can help solve regression tasks by using data that are more easily accessible to predict values of properties that are expensive or computationally intensive to compute. This is a common practice for scientific researchers, as seen by how K. Gong and coworkers successfully calculated binding energies using density functional theory (DFT) and then extrapolated them with second-order polynomials to avoid further DFT calculations on metal cations not previously calculated[7]. The main benefit of using machine learning models to make these predictions is that they can calculate predictions nearly instantaneously with no human input or background information on these properties.

Graph Neural Networks Used with Molecular Structure Data

The benefits of utilizing neural networks to estimate materials property data become apparent when considering interatomic potentials or inferring properties of materials that require DFT but are prohibitively expensive to obtain large samples of (e.g., high-entropy materials or catalysts). Graph neural networks specifically serve as relevant tools to predict these properties because of their ability to input the molecular structures of these high-entropy materials and predict properties rapidly. Such was the case when C. M. Clausen and coworkers successfully performed zero-shot inferences of adsorption energies of high-entropy materials utilizing a graph neural network. Here, zero-shot inference refers to the direct prediction of relaxed structure models and adsorption energies of materials using a machine learning model that was not explicitly trained to output those relaxed structures given an initial structure. Their model possessed inference speeds that were "practically instantaneous" thanks to the low number of parameters used in the neural network. Despite already achieving "state-of-the-art accuracy," yielding mean absolute errors in adsorption energies of 0.04 eV (adsorption energies of these high-entropy materials are normally on the order of magnitude of 10^0), with the fraction of the time DFT would take to emulate the high-entropy materials, C. M. Clausen and coworkers stated that a hyperparameter optimization could "further improve the performance"[17]. C. M. Clausen and coworkers' work highlights the possibilities of using machine learning to accurately estimate data on materials that were out of the domain of previous experiments or data collections. Optimizing hyperparameters and increasing the depth of the neural networks appear to be valuable steps to take in the future because of the amount of time and resources these machine-learning models can save when compared to running DFT calculations with supercomputers. Further, substituting machine learning models in place of DFT calculations allows research normally involving computational quantum mechanical modeling methods to be much more accessible to researchers around the world, as supercomputers for scientific research may not be readily accessible for researchers, especially those in less developed countries.

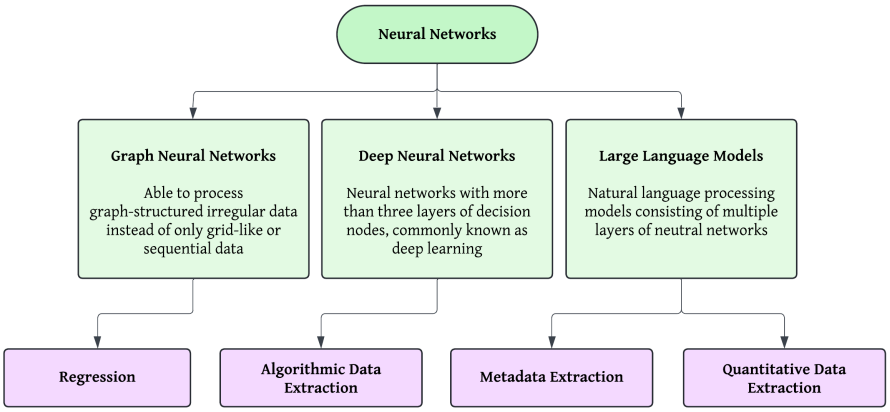

Figure 4: A Non-Comprehensive Diagram Depicting Types of Neural Networks and Tasks

Decision Trees Used with Continuous Data

Although the nature of decision trees makes them seem suitable for classification tasks related to categorial variables, some types of decision trees can be powerful tools that can predict quantitative variables rapidly. Similar to the application of graph neural networks, decision trees can be used to estimate properties of materials that are on par with the quality obtained from DFT calculations and experimental results. For example, K. Kaufmann and coworkers were able to use random forest decision trees to estimate the synthesizability of high-entropy materials [18]. K. Kaufmann and coworkers first chose a set of around 800 attributes of materials that may be relevant for synthesizability to feed as features for the machine learning model. Then, they utilized a feature named the SelectFromModel in Scikit-learn [19], a machine learning tool in Python, to narrow down the eight most relevant features to be used to train their model. Among these eight features related to chemistry, physics, or thermodynamics, K. Kaufmann and coworkers were able to predict which of these features contributed the most towards the final prediction of synthesizability [18]. This is a strong benefit of using decision trees for regression tasks: allowing us to peek inside the black box of machine learning models. Thanks to its flowchart-like structure, decision tree models are one of the most interpretable machine learning models available. In scenarios like these, where some sort of connection between quantitative variables is expected, being able to predict which quantitative attributes were responsible for the data can bring researchers one step closer to understanding the underlying cause and science that they were unaware of, which is why the interpretable quality of decision trees is valuable.

Unfortunately, some decision tree algorithms are much more complicated in nature and are harder to interpret than random forest decision trees or regular decision trees. Such is the case for extreme gradient boost and adaptive boost decision trees. Because normalizations and optimizations are applied to extreme gradient boost trees and because adaptive boost decision trees compare the training data, they are intrinsically too complex to interpret or understand. Still, these decision trees take comparatively much less time to make predictions. For example, when J. Deb and coworkers predicted the ablation performance of ceramic matrix composites using various decision trees to compare results, the extreme gradient boost and adaptive boost decision trees took significantly less simulation time [12]. Specifically, when using the performance parameters that resulted in the greatest R-squared scores and the least mean absolute errors among all decision trees, the extreme gradient boost and adaptive boost decision trees took 1 second and 2 seconds accordingly, the random forest decision tree took slightly longer than 2 minutes, the regular decision tree took longer than 18 minutes, and the gradient boost decision tree took close to 9 minutes [12]. This shows that the decision trees that are too complicated to be interpreted still serve a purpose, as they can drastically improve computing efficiency while still retaining high accuracy.

Flowchart of Machine Learning Methods to Utilize for Different Regression Tasks

The flowchart depicted in Figure 5 shows a summary of what machine learning models one should use depending on the type of regression task assigned. The properties of ceramics

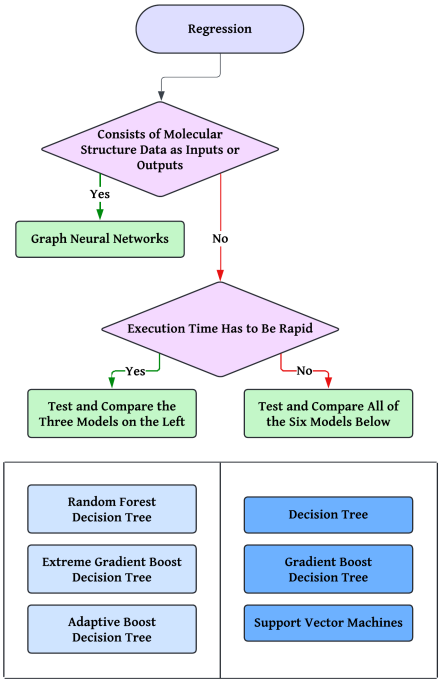

Figure 5: A Brief Overview of What Machine Learning Models to Use Given Different Scenarios

are dependent on the molecular structure of the material. As such, it is plausible that some form of molecular structure data is involved with the inputs or outputs of a given regression task. In such cases, it is encouraged to use a deep neural network to be able to take full advantage of the structured data. Further, predicting molecular structures of materials as an intermediate step before predicting a quantitative property could be useful, as knowing the molecular structures can help determine the distribution of electron density across the material, find distances between atoms, and find angles between bonds, all of which allow us to utilize DFT or quantum mechanical modeling to predict other properties of a material, if such level of accuracy is necessary for a project or experiment. For example, a graph neural network can be used to help narrow down which candidate has the significance to further test using expensive DFT calculations to make more accurate results.

When predicting continuous, quantitative data, you can utilize various decision trees or a support vector regression. In particular, the extreme gradient boost decision tree, adaptive boost decision tree, and random forest decision tree models execute the fastest among these six machine learning models shown in Figure 5. This was shown experimentally in J. Deb and coworkers' experiments [12], and aligns with the properties of the model: the extreme gradient boost tree is an optimization of a gradient boost tree that implements parallelization to speed up the tree-building process, while the adaptive boost decision tree and the random forest decision trees generate small trees to aggregate their results, which takes less time than generating a larger, more complex tree. Other than the apparent difference in the expected execution time of these models, it is hard to predict which model will perform the best in terms of accuracy or precision because of how unique these models are. Here, it is worth noting that the comparative performance of these models remains relatively constant even as the test data increases. Despite the small changes in the performances of the regression models as more of the test data was used for training, the relative order of performance among the regression models stayed constant in J. Deb and coworkers' experiments [12]. This signifies that a researcher can simulate these six for a smaller percentage of the test data, say 10%, determine which model suits their need the best, and then continue the training for that model only, saving time and energy.

Machine Learning Models Used for Data and Algorithmic Metadata Extraction

Machine learning models such as large language models and deep neural networks can help build databases by extracting data, metadata, and algorithmic data. The vast sea of literature and data floating online makes it overly time-consuming or expensive to manually extract information by hand. For large-scale projects with substantial budgets, manually extracting certain scientific data and algorithmic data would be more plausible. However, in the ceramics community, there hasn't been enough effort or budget to create a centralized and organized database hub for ceramics property data, as noted by S. Freiman and J. Rumble [9]. As it stands, it is beneficial to be able to generate databases with sufficiently high accuracy without requiring extensive manpower, and a combination of machine learning models can help researchers make those databases with minimal human input and without domain expertise.

Large Language Models for Data Extraction

Despite the immensely practical and applicable nature of large language models (LLMs), it may seem troublesome to use them to extract data, because with such a black-box system, one cannot be entirely certain that the output is reliable. However, when combined with human supervision, LLMs can extract data from research papers with impressive accuracy, as shown in the works of M. Polak and coworkers [10]. Below Figure 6 is the diagram of the method of data extraction proposed by M. Polak and coworkers.

In the proposed method, the text is initially preprocessed, a standard practice in data extraction methods. Then, a LLM is used to perform a zero-shot binary classification on the individual sentences to determine whether the sentence is relevant or not in collecting the data in question. Here, the "zero-shot" characteristic refers to the fact that the LLM used has never classified sentences in relation to the data in question beforehand. Because of this nature of the initial classification process, the accuracy of the binary classification can be lacking at this step of the procedure, and hence the human intervention in step two was introduced to improve the performance of the model. In step two, the 100 (or up to 200) "highest scoring" (most likely to be true positives) data from step one is manually, and therefore accurately, classified into relevant or irrelevant classes. Then, this training set created through human intervention is used to fine-tune the LLM, a procedure necessary to significantly improve the accuracy of the overall classification.

After the fine-tuned LLM reclassifies all the sentences, the user sorts the sentences based on the probability of the sentence being relevant, and traverses down the list until the desired recall is reached. To do so, the user can keep track of the precision of the sentences reviewed and use it to estimate the current recall. The authors suggest that to reach a recall of 90%, the user can extract the data from the sentences until the precision of the reviewed sentences reaches around 80%. The authors also suggest that a recall of 90% is a "reasonable value" to terminate this reviewing process because "the precision sharply drops for higher values, which diminishes returns for the human time involved" [10].

Although the procedure involves human input and is not fully automatic, the authors note that, based on two trials of

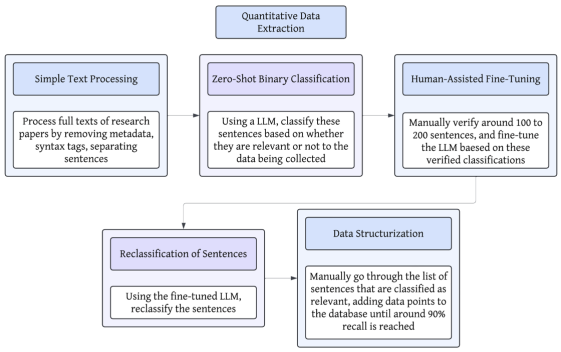

Figure 6: Depiction of Steps Necessary for NLP/LLM Data Extraction Proposed by M. Polak and Coworkers [10]

the procedure with 100 and 668 papers, respectively, "the LLM classification step (Step 1 in Figure 6) typically removes about 99% of irrelevant data and leaves only about 1% to be further analyzed," and therefore, human labor is "dramatically [reduced]" [10] compared to a manually curated database. Also, despite using a LLM to classify the sentences, a high precision is retained through the manual structurization process in the last step. Further, because the method is based on classifying sentences as relevant or not, the entire procedure can be applied to captions included in the papers, which can reveal relevant complex datasets such as tables and figures that contain the data in question [10]. Once the complex datasets are marked as relevant, the data can either be manually or automatically extracted using a preprogrammed algorithm, depending on the number of tables or figures and the feasibility of manually extracting the data.

Overall, the method's strength lies in the fact that the minimal human input involved (manually labelling 100 to 200 sentences during the fine-tuning step and structuring the data in the end) can result in a database with nearly perfect precision and high recall that is "comparable to a fully human curated database, but at 100 times or less the human effort" [10]. In fact, in a separate trial of the model aside from the bulk modulus dataset consisting of sentences across 100 randomly selected papers that was used as a benchmark, the authors were able to curate a database for critical cooling rates for metallic glasses after extracting data from 668 papers, an unreasonable number of papers for a human researcher to analyze manually. They resulted with 129 unique data points, which is larger than the "most state-of-the-art and complete" manually curated database of critical cooling points containing only 77 unique datapoints [10]. Not to mention, the resulting high precision and recall above 90% is noteworthy because even "state-of-the-art" named entity recognition-based tools such as the ChemDataExtractor2 achieved a 52% precision at 37% recall in the same scenario [10]. And finally, since the initial classification is carried out in a zero-shot fashion and the fine-tuning only after the zero-shot classification, the method can be applied to a variety of data types.

As the development of LLMs continues, using LLMs to extract data from research papers in a human-assisted manner seems to be a promising way to extract data rapidly (the proposed method by M. Polak and coworkers could create databases of up to around 1000 entries in a single workday [10]) and without needing a data scientist.

Deep Neural Networks for Algorithmic Data Extraction

To make the ceramics property databases adhere to the "interoperable" aspect of the FAIR principle, it is important to provide access to the code used to analyze the data present or provide a way to obtain such code without having to train domain experts in data science. Machine learning models, specifically deep neural networks (DNNs), can be used to extract algorithmic data so that the databases can contain pseudocode extracted alongside the data extracted using large language models. I. Safder and coworkers were able to create a procedure dedicated to extracting algorithmic data using clever techniques, showcased in Figure 7 [20].

As depicted in Figure 7, they first converted the PDF files into text files and looked for "sparse boxes" that were left sparse because of the formatting required for showcasing programming languages. Then, based on 60 features, including font, structure, and content of the text region, the DNN was trained to classify whether the text given contains algorithmic data or not. The classification resulting from this procedure had an F1-score of 93.32%, "outperforming the state-of-the-art techniques by 28%", and a near 80% accuracy [20].

Flowchart of Machine Learning Methods to Utilize

Figure 8 depicts a flowchart to help guide which machine learning models to use and how to use them to create databases through extraction. First, from resources with an abundance of research papers, researchers can select candidates that may be relevant to the material property in question. After selecting the candidate documents, the researchers can now use the

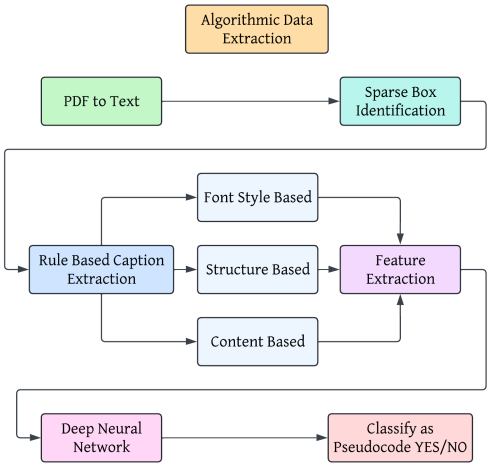

Figure 7: Depiction of Procedure for Extracting Algorithmic Data from Literature, Proposed by I. Safder and Coworkers[20]

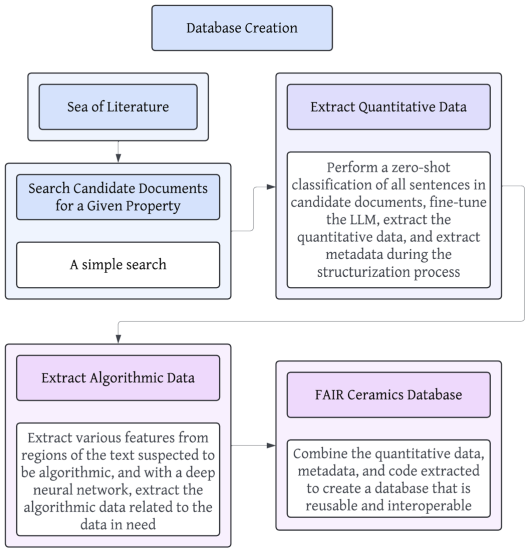

Figure 8: Proposed Method of Database Creation Utilizing Machine Learning Models

data extraction method involving fine-tuning LLMs through a modest number of manual classifications dipicted in Figure 6 to determine relevant sentences, figures, or tables that contain the data in question. During the structurization of this data, researchers can also obtain the metadata associated with the data. This can be done by mapping each of the individual sentences to the source, which can be used during the final structurization process to track the metadata related to the data value, as the researchers extract the data from sentences manually during this step. Finally, the researchers can extract any algorithmic data that can be used to analyze the data presented in various research papers, if any exists, to implement as resources for the database extracted. Overall, the resulting database would have the metadata required to decide whether the experiment or simulation used to obtain the data was reliable, and the algorithmic data to interoperate the data.

Discussion

Materials science and research can help develop alternative materials to replace carbon-intensive additives, enable a higher percentage of recycled content in ceramics, and develop kilns that are affordable and efficient at containing heat while being compatible with cleaner energy sources other than fossil fuels. However, specifically in the ceramic community, some challenges persist that can hinder these technological innovations. Applying the methods proposed earlier and utilizing various machine learning models can aid in the process of overcoming these following challenges and promote further materials innovations toward sustainable ceramics.

One of the biggest challenges the ceramic community must overcome for innovative research, as presented by E. Guire and coworkers, is the severe lack of infrastructure and a data-sharing platform. Currently, the ceramic community is given a minimal number of databases to research with, and more databases that adhere to the FAIR principles and promote data-sharing are necessary to accelerate data-driven research processes[21]. Using the proposal described in Figure 8, we can create structured databases with metadata to validate the integrity of the data included.

Additionally, machine learning models have the potential to extrapolate data outside the bounds of the data they were trained on. By training regression models such as the various decision trees, a graph neural network, or support vector regression models, researchers can extrapolate outwards from a set of training data to achieve predictions with incredible accuracy, following the proposal described in Figure 5. However, it is crucial to note that machine learning models are most accurate at making predictions in domains similar to those of the trained data set. Although making predictions outside the domain of the data set can be useful by providing interesting insights or acting as a rough benchmark for material selection, the extrapolated result should be approached with caution and shouldn't always be taken at face value.

Another challenge present in the ceramic society is the limited data accessibility and usability of ceramic materials[21]. We believe that the methods for creating databases discussed in this paper offer a reassuring prospect to help resolve this challenge. Through the methods proposed, large organizations such as the American Ceramic Society can create a hub for a comprehensive list of relevant ceramic materials data. This implies the possibility of a centralized database that one can access with a single access point and subscription or signup, reducing the redundancy of databases over the World Wide Web and improving the efficiency of researchers. Additionally, the algorithmic data extracted using the methodology proposed in Figure 8 allows ceramics researchers to analyze and utilize the data in the databases to their fullest potential without having to rely heavily on data scientists. This will reduce the gap between the ceramic researchers and the data scientists, allowing the researchers to acquire the code necessary on the spot, accelerating the research process. Further, if no algorithmic data were to be present, the centralized database hub could implement a virtual assistant based on LLMs that can generate the code needed for the user on demand, just like how ChatGPT can generate functioning code snippets upon natural language instructions.

Another significant challenge the ceramic community must overcome is the lack of appropriate computational approaches to accurately calculate large-scale models of ceramics[21]. Machine learning is one of the keys to solving this challenge. By training machine learning models on high-fidelity simulation data to build accurate and universal interatomic potentials, they can then be used to simulate larger systems at a fraction of the computing time required by DFT. Some commercial platforms have already broken ground in this area in catalysis, organic chemistry, and polymers, but their application to ceramics remains limited. Considering how graph neural networks can input and predict excited and ground state molecular structures of materials accurately at a fraction of the time of a DFT simulation, further research into machine learning and their applications to ceramics has the potential for scientific breakthroughs in the future.

Aside from challenges specific to the ceramic community, there are certain issues that must be considered when either applying the methods proposed or when handling databases in general. First, there may be missing or incomplete data when accessing papers or organizing databases. It would be most accurate to contact the creator of the data who conducted the experiment, simulation, or regression, but this may not be possible, and some form of data imputation may be necessary. If such an issue arises, we advise the reader to explore reviews and studies that compare deep learning and conventional methods for missing data imputation[22], analyze missing value imputation methods carried out in the past[23], introduce and assess machine

learning-based imputation methods[24], or provide useful rules and instructions for imputation[25].

Another issue to consider is the standardization of data. For any data structurization process, it is crucial that the data is well-standardized and that all graphical data are properly converted into tabular data to ensure that the resulting database adheres to FAIR principles. In the database creation method proposed in Figure 8, the final structurization process is carried out manually. Therefore, it is of utmost importance that the researchers who carry out this finalization process are aware of the standards of the database and that the data manually extracted from any figures or tables conform to those standards.

When creating databases, graphical data (e.g., figures, tables) are important sources of data. One way to improve the database creation process is to utilize accurate tools that can read these forms of data and convert them into entries up to the standards of databases. Even though some tools exist (e.g., Engauge Digitizer[26]), it is important to have more developed and easy-to-use algorithms and tools that process graphical data into tabular data.

Finally, we would like to caution the readers that the recommendations presented here are based on a limited literature survey and are yet to be proven in practice. They should be taken as a first step towards addressing challenges facing the research for sustainable ceramics. Additionally, sources like The Materials Project (https://next-gen.materialsproject.org/), a project dedicated to pre-compute properties of materials using high-throughput computing, can offer valuable materials data that can be utilized for comparing with experimental data or creating datasets to train machine learning models[27].

Methods

We have mainly used the Google Scholars search engine to search prompts that relate various machine learning models (e.g., random forest decision tree), tasks (e.g., regression), or programs (e.g., Scikit-learn) to ceramics or materials research to find research papers, journals, and articles relevant to this paper. To verify the relevance of sources, especially if some time has passed since publication, we've checked the publication date, which is of the greatest importance for the relevancy of machine learning methods for contemporary database creation, and how relevant the proposed methods are for the purpose of creating ceramic property databases. To verify the relevance of challenges present in the ceramics community, we've checked large databases such as the National Institute of Standards and Technology, and to assess the quality of the methods proposed, we've checked the design of the study, such as what type of data (e.g. experimental, DFT calculated) was used to train the machine learning models and the statistical measures provided. Also, we have visited websites created by reputable sources, such as IBM, to obtain general information regarding machine learning models. To organize the information obtained from surveying these sources, we've categorized these sources based on the machine learning task performed and the model used, which aligns with the overall structure of this paper as well.

Acknowledgements

Despite the use of the 'authorial we', common in academia, this paper is the sole work of its author. Additionally, I would like to thank Dr. Ahmed Elnabawy (Cairo University, Giza, Egypt) for advising me and giving me invaluable support throughout the research and writing process.

References

1 European Ceramic Industry Association, *Ceramic Roadmap to 2050*, The European Ceramic Industry Association, Brussels, Belgium, 2021.

2 D. D. F. D. Rio, B. K. Sovacool, A. M. Foley, S. Griffiths, M. Bazilian, J. Kim and D. Rooney, *Renewable and Sustainable Energy Reviews*, 2022, **157**, 112081.

3 Z. Utlu, A. Hepbasli and M. Turan, *Drying Technology*, 2011, **29**, 1792–1813.

4 E. F. S. Ciacco, J. R. Rocha and A. R. Coutinho, *Applied Thermal Engineering*, 2017, **113**, 1283–1289.

5 A. Mezquita, E. Monfort and V. Zaera, *Boletin de la Sociedad Espanola de Ceramica y Vidrio*, 2009, **48**, 211–222.

6 M. S. Asl, B. Nayebi, Z. Ahmadi, M. J. Zamharir and M. Shokouhimehr, *Ceramics International*, 2018, **44**, 7334–7348.

7 K. Gong, K. Yang and C. E. White, *Frontiers in Materials*, 2023, **10**, year.

8 S. Kim, M. Lee, C. Hong, Y. Yoon, H. An, D. Lee, W. Jeong, D. Yoo, Y. Kang, Y. Youn and S. Han, *Scientific Data*, 2020, **7**, 387.

9 S. Freiman and J. Rumble, *American Ceramic Society Bulletin*, 2013, **92**, 34–39.

10 M. Polak, S. Modi, A. Latosinska, J. Zhang, C. Wang, S. Wang, A. Hazra and D. Morgan, *Preprint at*, 2023.

11 IBM, *What are support vector machines (SVMs)?*, 2023, `https://www.ibm.com/topics/support-vector-machine`.

12 J. Deb, J. Gou, H. Song and C. Maiti, *Journal of Composites Science*, 2024, **8**, 96.

13 IBM, *What is a decision tree?*, 2023, `https://www.ibm.com/topics/decision-trees`.

14 IBM, *What is random forest?*, 2023, `https://www.ibm.com/topics/random-forest`.

15 IBM, *What is a neural network?*, 2023, `https://www.ibm.com/topics/neural-networks`.

16 IBM, *What are large language models (LLMs)?*, 2023, `https://www.ibm.com/topics/large-language-models`.

17 C. M. Clausen, J. Rossmeisi and Z. Ulissi, *Preprint at*, 2024.

18 K. Kaufmann, D. Maryanovsky, W. Mellor, C. Zhu, A. Rosengarten, T. Harrington, C. Oses, C. Toher, S. Curtarolo and K. Vecchio, *NPJ Computational Materials*, 2020, **6**, 42.

19 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay and G. Louppe, *Journal of Machine Learning Research*, 2012, **12**, year.

20 I. Safder, S. Hassan, A. Visvizi, T. Noraset, R. Nawaz and S. Tuarob, *Information Processing and Management*, 2020, **57**, 102269.

21 E. Guire, L. Bartolo, R. Brindle, R. Devanathan, E. Dickey, J. Fessler, R. French, U. Fotheringham, M. Harmer, E. Lara-Curzio, S. Lichtner, E. Maillet, J. Mauro, M. Mecklenborg, B. Meredi, K. Rajan, J. Rickman, S. Sinnott, C. Spahr and R. Weber, *Journal of the American Ceramic Society*, 2019, **102**, year.

22 Y. Sun, J. Li, Y. Xu, T. Zhang and X. Wang, *Expert Systems with Applications*, 2023, **227**, 120201.

23 W.-C. Lin and C.-F. Tsai, *Artificial Intelligence Review*, 2020, **53**, 1487–1509.

24 T. Shadbahr, M. Roberts, J. Stanczuk, J. Gilbey, P. Teare, S. Dittmer, M. Thorpe, R. V. Torné, E. Sala, P. Lió, M. Patel, J. Preller, A.-C. Collaboration, J. H. F. Rudd, T. Mirtti, A. S. Rannikko, J. A. D. Aston, J. Tang and C.-B. Schönlieb, *Communications Medicine*, 2023, **3**, 139.

25 L. Ren, T. Wang, A. S. Seklouli, H. Zhang and A. Bouras, *Information Systems*, 2023, **119**, 102268.

26 C. Young, *Converting graphs to tables of data*, 2024, `https://engineerexcel.com/converting-graphs-to-tables-of-data/`.

27 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *APL Materials*, 2013, **1**, 011002.