A Theoretical Review on Improving Predictive Accuracy and Mitigating Overfitting in Materials Informatics

Razak bin Osman
Comilla University, 3506 Comilla Univ Rd, Kotbari 3506, Bangladesh.
binosman6788@gmail.com

Correspondence should be addressed to Razak bin Osman : binosman6788@gmail.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/JCIMS202402007.
Received 29 December 2023; Revised from 28 March 2024; Accepted 04 April 2024.
Available online 29 April 2024.

Abstract – Within the field of machine learning, where computers are required to identify the best match for a particular set of data, overfitting is a typical concern. To effectively enhance the accuracy of prediction, this research seeks to investigate the problem of overfitting in the field of machine learning, and recommends novel techniques to establish hypothesis functions retrieved from data. This paper also discusses the necessity for more data collection, and potential challenges related to handling unrepresented datasets. To effectively predict data, this paper puts more emphasis on the selection of effective descriptors and feature extraction elements, with major focus on entropy-based and decision tree models within the field of informatics. In addition, this paper reviews principal component analysis (PCA) and model interpretability of applications. To enhance performance, this research ends with a discussion on the selection of standard models, and machine learning algorithms. The discussions in this article provides a basis of understanding the processes involved in the advancement of content-based reporting models, emphasizing the necessity of gathering essential data, developing sophisticated models, advancing them, and putting them to practical applications.

Keywords – Overfitting, Material Informatics, Principal Component Analysis, Predictive Accuracy, Descriptor Selector.

I. INTRODUCTION

The field of Materials Information (MI), as depicted in [1], represents the concept of materials engineering, which employs information science to enhance materials applications, analysis, selection, and understanding. With the aim of effectively and consistently gathering, arranging, evaluating, and disseminating a broad spectrum of product data, the area is still in its infancy. Materials information (MI), which is heavily reliant on and goes beyond conventional representations of interrelationships, aims to decrease the time and uncertainty involved in the development and deployment of innovative goods. Combinatorial chemistry, process modeling, product life cycle management, materials data management, and materials databases are some further, more specific definitions of the topic. MI goes above and beyond these guidelines by applying the understanding that is obtained from data gathered on one material to other materials in order to get more profound insights and understanding. The relevance of every single data item may be greatly increased by gathering pertinent information.

Scientists are turning more and more to machine learning—a discipline concerned with the analysis and development of algorithms that can learn from data and make predictions without direct human intervention—to accomplish these goals. In order to engage in machine learning, it is essential to have historical data, regardless of the individual issue being investigated. Therefore, it is necessary to have access to clean, curated, and trustworthy data that is relevant to the issue being investigated. If this data is not currently accessible, then a proactive effort must be made to create it. A such data collection may consist of a list of different materials belonging to a certain chemical class, together with a corresponding measured or calculated attribute of those materials (see to **Fig 1a**).

In the field of ML, the term "input" is used to refer to the material being analyzed, whereas the term "target" or "output" is used to refer to the specific attribute of interest [2]. The definition of a learning issue (see **Fig 1b**) is as follows: When provided with a data set including information on the properties of materials, what is the most accurate prediction for the property of a new material (NM) that is not included in the original data set (ODS)? If there are enough instances available in a suitably large data set, and if the NM belongs to the same chemo-structural class as the materials in the ODS, we anticipate that it should be feasible to construct such an estimate. It is desirable to include uncertainty in the forecast, since this may shows whether the new example falls within or beyond the range of ODS.

In the field of materials science, scholars have employed the Least Absolute Shrinkage and Selection Operator (LASSO) [3] technique to create power series developments, such as partition function cluster expansions for alloys. This approach has shown to be much quicker, by many orders of magnitude, compared to previous methods like genetic algorithms. Tree-based models are used for the purpose of optimizing the density of 3D printed parts, choosing dopants for ceria water splitting, and forecasting defects in steel plates. Clustering, in conjunction with principal component analysis (PCA), has proven effective in simplifying complex, multidimensional microscopy data into meaningful local structure representations [4]. Machine learning techniques in materials informatics have several potential applications, but they also carry the danger of misapplication and misinterpretation if used without proper safeguards or adapted from other data issues.

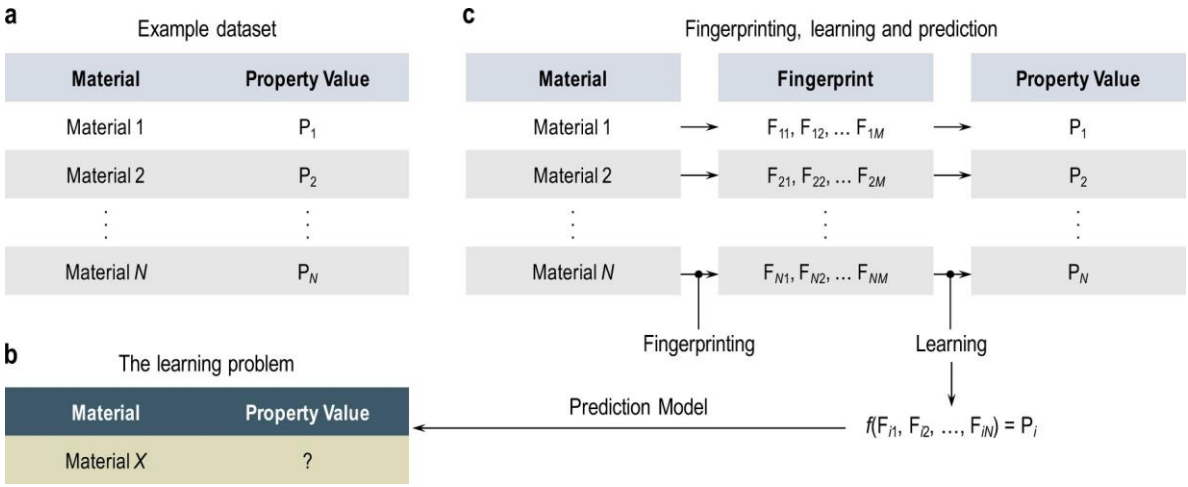

Fig 1. The fundamental components of ML in the sector of materials science. Schematic representation of a sample dataset, articulation of the learning issue, and generation of a surrogate prediction framework via the learning and fingerprinting procedures. M and N represent the quantity of samples of training and the quantity of fingerprint (or feature or descriptor) components.

The study explores the issue of overfitting in machine learning and its influence on materials informatics. This highlights the need for alternative methods for deriving hypothesis functions directly from data and the limitations of heuristic approaches. The study examines the challenges associated with the choices in content informatics models, translation models, and application of regularization techniques The study highlights the importance of expert knowledge in a field and the importance of consideration well when selecting model and algorithm. The study provides valuable insights into the content informatics model building process and proposes strategies to increase model accuracy and performance. The rest of the paper has been arranged as follows: Section II presents a discussion of overfitting, descriptor selection, features extraction, model interpretability and PCA in materials science. Section III focusses on critically defining cross-validation and regularization aspect. Section IV discusses the design of materials informatics models. Lastly, Section V presents conclusions to the research.

II. BASIC CONCEPTS IN MATERIALS SCIENCE

Overfitting

Learning algorithms (LA) primarily operate by exploring a set of functions, often referred to as the hypothesis class, in order to find a function that accurately matches the provided data. Due to the vast number of functions involved, it is not feasible to examine each individual hypothesis function. Therefore, an alternative approach is required to create the hypothesis functions directly from the data. Formulating this search often involves specifying an objective function, such as the count of erroneously predicted data points, and using different methods to identify a function that decreases this objective function (OF), which is known to be NP-hard. As an example, solving the challenge of determining the smallest decision tree (DT) or weights of a neural network (NN) are both considered NP-complete issues. Therefore, heuristic techniques, such as greedy search (GS) for DT and gradient descent (GD) for NN, have been effectively used.

Naturally, the inherent inefficiency of these heuristic algorithms immediately suggests a logical avenue for further investigation: develop algorithms that are capable of more effectively exploring the hypothesis class. Consequently, there has been a significant amount of research focused on using second order techniques to train NN and doing more comprehensive searches in the process of learning rule sets and DT. Paradoxically, during the evaluation of these algorithms on actual datasets, it was discovered that their presentation often proved to be worse to that of simpler methods such as simulated annealing and GS or GD. In general, it is more advisable to refrain from optimizing. An important tendency in ML research has been the development and cultivation of connections across previously separate domains, such as statistics, connectionist learning, symbolic learning, and computational learning theory,. The use of statistical analysis was important in resolving this paradox. The primary issue stems from the configuration of the ML duty.

Learning algorithms (LA) undergo training using a specific collection of data, and then utilizes this training to generate predictions on novel data points (DP). The objective is to optimize its predicting new DP accuracy, rather than just focusing on its accuracy on the training data (TD). Excessive efforts to achieve an optimal fit to the training data may result in the inclusion of irrelevant data noise, as the focus shifts towards remembering specific idiosyncrasies of the TD rather than discovering a universally applicable prediction rule. It's common to refer to this situation as "overfitting." As a result, the optimization problem's stated objective function, "minimize error on the TD," is in fact false. Many studies have tackled this issue by enhancing the objective function by the inclusion of different penalty factors, such as generalized cross-validation (CV) regularization techniques, minimum description length algorithms , and others. These phrases aim to forecast the accuracy of a model when applied to data outside of the training set, based on its accuracy inside the training set.

By rectifying the OF, we may once again use our arsenal of optimization algorithms to address this novel optimization issue (although with the potential for equivalent or worse computational complexity issues). Empirical findings support the resolution of this issue. However, instead of trying to solve this modified OF, we may consider the basic GD and greedy algorithms (GA) as implicitly including corrective terms [5]. Put simply, a GA may be considered minimal when its goal is to discover the smallest DT tree that accurately represents the data. However, it can be regarded as superlative if its aim is to identify a tree that decreases a combination of a penalty term and DT size that accounts for the disparity between the TD and the final test information. In the field of ML, it is advantageous to deliberately choose inferior solutions. This is fortunate, since the initial optimization difficulties were insoluble. Ultimately, we possess a polynomial-time algorithm that effectively does the desired task. Through the concept of “under computing,” we prevent the occurrence of “overfitting.”

Underrepresented Classes

Another nuanced but significant aspect is to the inadequate representation of certain categories in data during categorization tasks. Discovering unique properties that increase thermoelectric power, such large ZT or high TC superconductivity, is a common goal for materials scientists. Suppose there exist many perovskite compositions with the general stoichiometry ABX3, where X denotes an anion and A and B represent cations. It is estimated that maybe fewer than 5% of these compositions exhibit superconductivity above 60K (although this is a speculative approximation). If the primary objective of your framework is to optimize the overall accuracy classification, the ML algorithm will achieve a high level of performance (95% accuracy) by assuming that no material can ever exhibit high TC superconductivity.

Accurately distinguishing the 5% as high temperature superconductors has more significance than potentially mislabeling the other 95% [6]. Fortunately, ML practitioners have long been addressing these challenges and there are methods to alleviate the issue. The techniques primarily emphasize resampling approaches to balance the dataset and incorporating balances into current learning algorithms to address misclassification of the minority class. An often-overlooked solution for data scientists is to gather more data, a task that may be effectively accomplished by a materials scientist.

Overfitting Example

Decision tree, a ML approach, are susceptible to such issues. We will now use decision trees as an example to go into the intricacies more comprehensively. DT function by iteratively dividing the information using a sequence of rules and inducers (refer to **Table 1**) established based on an attribute value examination. The final outcome is comparable to a flow chart consisting of hierarchical rule nodes that lead to various forecasts. When attempting to anticipate whether metal oxides exhibit Mott insulator behavior, a rule may be established that asserts any materials having an optical band gap (OBG) [7] less than 0.01 eV are not Mott insulators (MIs) [8]. Other criteria might be implemented, specifying that substances with an odd electrons number in every unit cell and an OBG greater than 0 eV are classified as MIs. Nodes that occur higher up in the tree partition a greater number of samples than those further down, and might be considered more significant in the stratification process.

Table 1. Additional Decision Tree Inducers

Algorithm	**Description**
TI	A decision tree with one level that uses a single characteristic to classify occurrences. Missing people are given "a special value." Provide support for both discrete and continuous characteristics.
LMDT	The components of a multivariate decision-based tree test consist of linear combinations of the qualities.
PUBLIC	uses MDL cost to integrate pruning and growth in order to minimize computational complexity.
FACT	A previous iteration of QUEST. selects a feature for separating each node using statistical testing, and then discriminant analysis is used to locate the split point.
MARS	The tensor products of linear spines are used to simulate a multiple regression function.
CAL5	Created especially for properties with numerical values.

New users of machine learning may not fully realize that by altering the criteria for picking partitions, they might notice distinct qualitative outcomes across decision trees. As an example, we used data, which focused on forecasting high-temperature piezoelectric perovskites. Our analysis confirms that there is a possibility of obtaining outcomes that vary in nature. When the gini impurity attribute is used, both the global instability index (GII) [9] and the computed Goldschmidt tolerance factor (TF) (emerge as significant features. In the second scenario, where data or entropy gain is used instead, the

TF and GII are not considered as variables. However, other factors emerge, such as the disparity in ionic radii ratios between the A, B, and X = O ions, expressed as $\left[\left(\frac{rA}{rO}\right) - \left(\frac{rB}{rO}\right)\right]$. and the ionic character of the A-site.

Domain knowledge indicates that the TF, difference in A-site (AS) ionicity and ionic radii ratios are directly related to the size of the AS cation. These factors are believed to arise from the ionic solids' preferences for tight packing. The GII relies on the disparity between theoretical and computed bond valences, enabling the inclusion of local bonding influences with preferences of steric packing. If the GII can accurately forecast the stability of cubic perovskite oxide [10] compared to binary oxide [11] phases it may disintegrate into, it would allow for screening a larger number of compounds than if solely radius-based parameters were used for prediction. There are indeed several cations that have established bond valence characteristics but do not have the required tabular 12-coordinate ionic radii, which are essential for the TF calculation. Furthermore, the inclusion of the GII in the framework l might provide further insights into the bonding qualities that contribute to phase stability.

These trees were trained using identical information, implying that they should represent the same underlying physics. However, this assumption is not totally accurate. Constructing a tree that is very accurate in its representation requires significant computer resources. As a result, heuristic methods are used as an alternative, however they do not provide a guarantee of finding an optimum solution that encompasses the whole problem space. However, the resultant tree might differ even across several iterations of the same procedure. Expanding the permissible node layers number might potentially lead to convergence, but it can also result in a complex decision tree that is difficult to interpret. Here, the starting measure was 92 (±8%) with four nodes. However, after rebuilding the entropy-based model using 10-fold cross-validation and a high depth of seven nodes, we obtain a 95% accuracy confidence interval of 93 (±12%). The significant degree of precision in both instances suggests that a small number of structural characteristics, such as ratios of ionic radii and the optimum spacing between A and O atoms, are appropriate for determining whether an ABO3 composition would adopt the perovskite structure.

The accuracy drops to about 90% for both node amounts when the number of divisions is changed from 10 to 5, which is an intriguing aside [12]. Specifically, altering the ratio of training data from 90% to 80% results in a significant decrease in average accuracy. The significant fluctuation in performance resulting from altering the data partitioning suggests that the model is very responsive to new data. Consequently, we may anticipate that the model will not be as accurate in predicting the formability of novel compounds as first indicated by cross-validation. It is often unwise to assume that default hyperparameters, like folds number, are always ideal. Additionally, relying only on cross-validation may not provide an absolute truth or a definitive answer. Cross-validation is a method that relies on the assumption that the data provided adequately represents the whole population being studied. We are not asserting that one tree is unequivocally incorrect and another is unequivocally correct, but rather that each particular framework described in the literature is a product of several decisions made by the modeler. Domain expertise should be used to assess the effectiveness of a model, together with the indicated margins of error.

Descriptor Selection

Materials informatics deals with physically significant factors. The use of descriptors to characterize material characteristics is crucial for establishing accurate predictions and developing a comprehensive knowledge of relevant systems. Schultz and Cronin [13] have established several characteristics that define a high-quality descriptor. Specifically, a suitable description should be easier to ascertain than the actual trait, regardless of whether it is gained via computational means or experimental measurements. In addition, a minimum thickness must give the product a distinctive character. In materials science, explanations come from a wide variety of complexities. The periodic table provides information about atomic radius, atomic number, electronegativity, elemental period, or group, which can be used to make predictions about the type of system. The free energy of a molecule may be determined by using density functional theory and then linked to the stability of its phase. Experimental measurements may be conducted to determine densities and structural factors, which can then be used to predict mechanical properties. Furthermore, it is important to note that descriptors may also be formed by combining quantities from either the same or other levels.

Nevertheless, there is no commonly recognized approach for selecting descriptors. The selection of descriptors will mostly rely on the nature of the phenomenon under investigation. Dietterich and Michalski [14] offer a detailed explanation of compound descriptors that are produced from elemental and structural representations and meet the specified constraints. These characteristics may be attributed to both crystalline systems and molecular systems. The authors discuss the technique for generating descriptors for compounds in a schematic manner. The compound is comprised of atoms, each characterized by certain element kinds and neighboring surroundings influenced by other atoms. Each atom is characterized by a combination of elemental representations $N_{x,ele}$ and structural representations $N_{x,st}$, denoted as $N_x = N_{x,st} + N_{x,ele}$. Hence, compound ξ may be represented by a matrix of dimensions $N_a^{,\xi}$,N_x, where $N_a^{,\xi}$ is the number of atoms in the unit cell of compound ξ. The matrix that represents compound ξ, denoted as $X^{(\xi)}$, is expressed as:

$$X^{(\xi)} = \begin{bmatrix} x_1^{(\xi,1)} & x_2^{(\xi,1)} & \cdots & x_N^{(\xi,1)} \\ x_1^{(\xi,2)} & x_2^{(\xi,2)} & \cdots & x_{N_x}^{(\xi^x,2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{\left(\xi,N_a^{(\xi)}\right)} & x_2^{\left(\xi,N_a^{(\xi)}\right)} & \cdots & x_{N_x}^{\left(\xi,N_a^{(\xi)}\right)} \end{bmatrix} \quad (1)$$

where $x_n^{(\xi,i)}$ represents the nth representation of atom i in compound ξ.

There has to be a technique to convert the representation matrix into a collection of descriptors in order to compare various compounds, since it is merely a representation of the unit cell of compound π. One way to imagine this transformation is as a dispersion of data points in a N_x-dimensional space, and the representation matrix is just that. Mean, standard deviation (SD), skewness, kurtosis, and covariance are some representative variables that are used to describe the distributions and allow for comparisons between them. The interplay between the element type and crystal structure may be taken into account when the covariance is included.

Feature Extraction

Occasionally, it is not feasible to construct the optimal model using the originally chosen characteristics. An illustrative instance may include using regression analysis to forecast the activation energy based on observed diffusion data. Utilizing the logarithm of diffusion constants provides a more accurate alignment compared to fitting the unprocessed data. One may identify several physically realistic main descriptors based on the material system of interest and then create additional descriptors from them using a specific method. Possible approaches may include using dimensional analysis groups, basic relational characteristics,

PCA, or other techniques. It is important to use caution in all situations to prevent incompatible actions (for example, refrain from adding an ionization potential to an atomic radius). Edges are often retrieved as properties from the basic pixel information in picture data for the purpose of learning. Eureqa, a commercial program, enables the rapid generation of function sets, including Gaussian and exponential functions, among others. After extracting the features, it may be necessary to do down selection in order to test just the most significant characteristics. Gao and Li [15] utilized LASSO-based compressive sensing in conjunction with feature extraction to efficiently produce descriptors for the energy difference between the rock salt structure and zinc blende or wurtzite for 68 octet binary mixtures.

Model Interpretability

Scientists for materials aim to establish definitive causal relationships between the structure of materials, which encompasses many length scales, and their characteristics. When assessing a model used by Netflix, its performance is often measured based on its ability to accurately anticipate outcomes and its speed [16]. However, scientific models have additional requirements, including the need to have a small amount of frameworks and to comply with physical laws. If a framework cannot be effectively conveyed except amongst computers, its impact will be insignificant. The modeler has a responsibility to convert the outcomes of their labor into information that can be used by other materials scientists to assist in the process of discovering or implementing materials. However, manually removing factors to create a comprehensible model is often unfeasible. There are many beneficial strategies accessible in this scenario.

Principal Component Analysis

Principal component analysis [17] decreases the intricacy of high-dimensional information while preserving underlying patterns and trends. This is achieved by the process of reducing the data into a lower number of dimensions, which serve as condensed representations of the characteristics. High-dimensional data are often encountered in the field of biology and occur when various characteristics, such as the utterance levels of numerous genes, are assessed for each sample. PCA addresses two main issues associated with this sort of data: the high computing cost and the elevated error rate resulting from multiple test correction when assessing the link between each feature and an outcome. PCA is an unsupervised learning technique that has similarities with clustering. It identifies patterns in data without relying on any previous information about whether the samples belong to exhibit phenotypic differences or distinct treatment groups.

Principal component analysis (PCA) is a technique that lowers data by forecasting them onto smaller dimensions known as principal components (PCs). The objective is to identify the most effective information summary using a minimal PC. The initial PCs is selected to minimize the overall distance between the data points and their projection onto the PC, as shown in **Fig 2a**. By reducing this distance, we simultaneously increase the projected spots variance, σ2 (see **Fig 2b**). The selection process for the second (and subsequent) PCs follows a similar method, but with the added condition that they must be uncorrelated with all previously chosen PCs. As an example, the forecast onto PC1 is not associated with the forecast onto PC2, and we may consider the PCs as symmetrically perpendicular. This condition of no connection implies that the greatest number of PCs that may be obtained is either the number of features or number of samples.

The method of selecting PCs aims to maximize the connection (r2) between the data and their projection. This procedure is comparable to doing repeated linear regression (LR) on the information projected against each original data variable. For

instance, the highest coefficient of determination (r2) is achieved when the projection onto PC2 is used in multiple regression alongside PC1.

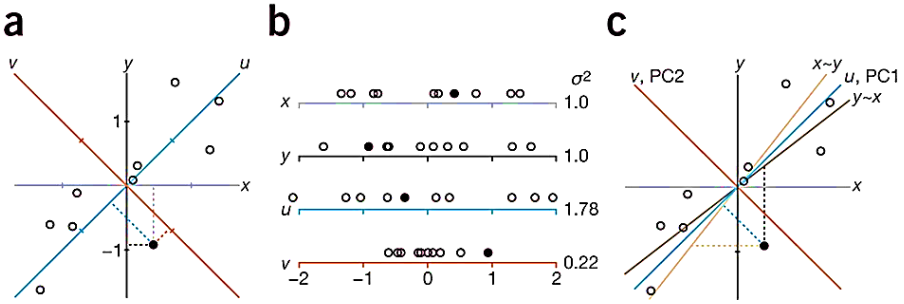

Fig 2. PCA maps information onto a lower-dimensional space in a geometric manner

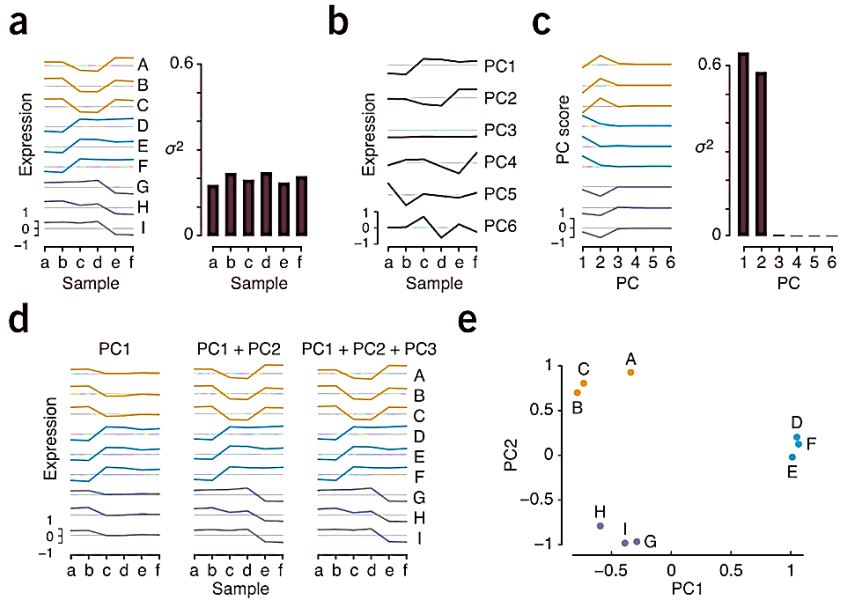

Fig 3. The PCA decrease of nine profiles of expression, reducing them from six dimensions to 2D

Principal component analysis is a potent method used to reduce the dimensionality of data (refer to **Fig 3**). Essentially, PCA involves transforming your data into a new coordinate system where the new axes, known as principle components (PCs), are formed by combining the original variables in a linear manner. The selection of each principal component is based on its alignment with the direction of maximum variance, while maintaining independence from other principal components.

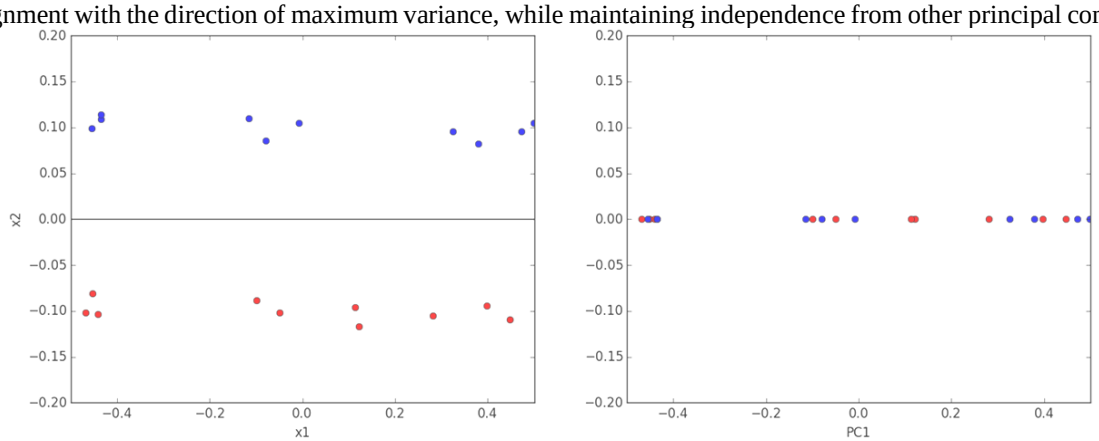

Fig 4. (Left), the dataset has two classes that are linearly separable, specifically designed for toy classification purposes. The optimal subspace generated by Principal Component Analysis indicated in black. Correct Mapping the data into this subspace eliminates the inherent distinctiveness of the original data.

When the data is adjusted to have an average of zero, these principal components serve as the eigenvectors for the covariance matrix of the samples. In principle, to accurately describe the original data, it would need the same number of PCs as there are original features. However, in practice, a subset of PCs is usually chosen based on a predetermined threshold

of explained variance or correlation with a certain characteristic. By using a few hundred PCs, it is possible to create a concise and accurate representation of a dataset including thousands of variables.

In **Fig 3**, the expression outlines of nine genes (A–I) in six samples (a–f) are represented by colors based on their similarity in form. The variance of expression of each instance is also shown. The profiles in a consist of PC1 to PC6. PC1 and PC2 exhibit prominent patterns, whereas the other PCs merely record minor variations. The transformed profiles are represented as principal component (PC) scores and the variance (σ2) of each factor value. The outlines were rebuilt using the first three principal components (PC1-PC3). The coordinates of 2D of each profile are determined by the scores of the first two principal components (PCs). However, it is important to exercise caution and conduct Principal Component Analysis (PCA) carefully. Authors recognize that interpreting PCs, particularly with image data, may be challenging. PCA does not provide a guarantee for the separation of data clusters. In fact, excessive projection onto principal components might result in the merging of information classes that were previously distinguishable (see **Fig 4**).

The structures of orthorhombic P_{nma} of the rare earth vanadate oxide perovskites ($La, RVO3, Ce, R = Y, Nd, Yb, Pr, Ho, Gd, Dy$) may be measured using symmetry-adapted models of distortion. Every mode inside the Pnma structure signifies a unique pattern of atomic dispersions from the ideal cubic phase. These patterns include first-order Jahn-Teller deformations of the VO6 octahedra ($M + 2M2 +$), in-phase VO6 octahedral rotations ($M + 3M3 +$), antipolar R-cation dispersions ($X + 5X5 +$ and $R + 5R5 +$), and out-of-phase octahedral turnings ($R + 4R4 +$). Subsequently, the net atomic dispersions associated with each mode, necessary to achieve the observed structure, are determined by calculating the size of the structural decomposition to calculate root-sum-squared dispersion in angstroms. Subsequently, it is possible to examine the correlations between these mode amplitudes and macroscopic observables, such as the temperatures associated with orbital ordering and magnetic spin. This analysis aims to determine whether there is a connection between electronic and magnetic transitions and the atomic structure.

III. CROSS-VALIDATION AND REGULARIZATION

Framework formalization involves incorporating a modifiable penalty on the size of model parameters into the cost function that is being reduced. Formalization is primarily used to mitigate overfitting in LR. It is widely known as ridge regression in this context, which lowers the parameters sum of squares, or LASSO in this context, which leads to absolute values sum decreases off the fit parameters. Regrettably, there is currently no known method for directly reducing the non-zero adjuvants number in an effective manner. However, these approximations often provide satisfactory results. The formalization penalty weight is incrementally raised, often in powers of 10, and evaluated using CV until the lowest test error is achieved. With the rise in penalty, an increasing number of model adjuvants are pushed towards 0. The larger an adjuvant is non-zero, the more its significance becomes in determining the performance of the framework.

One limitation is that techniques like RR and LASSO may eliminate associated features in a rather arbitrary manner. If feature A predicts both target C and feature B, the model may exhibit B predicting C due to regularization seeing A as redundant data. For example, the bond energy (BE) might be associated with the elastic modulus and melting point (MP). However, when attempting to forecast the model, the LASSO algorithm would disregard the BE in favor of the MP, even if there is no direct causal relationship. One may use a brute force approach to verify the presence of covariance across features. However, this method may become impractical when dealing with extensive feature parameters or sizes in the descriptor. Yong-Li and Yang [18] have thus modified Cross-Validation (CV) techniques to include random selection of both samples and characteristics. Ensemble approaches, like random forests, operate by aggregating several simple estimators, such as parameters average, and DT. Hyperparameters, like the simple learners number in an ensemble, the weight of the regularization penalty, and the proportion of samples/features to randomly employ, should be systematically tested in a cross-validation procedure. This is done to ensure that the final predictor is as basic as possible without being too simplistic.

In relation to the above statement, the selection of a machine learning model may significantly influence its comprehensibility for humans. Regressions and their regularized equivalents possess coefficients that indicate the magnitude of the impact of adjusting an input on the result. DT possess a structure like a flowchart, which is very comprehensible. Bayesian approaches enable the explicit incorporation of previous information. However, several methods like kernel ridge regression and kernelized support-vector machines use non-linear modifications that obscure the connection between the target variable and the original features. The intricate interconnections between nodes in neural networks sometimes make it difficult to provide a straightforward description of the machine's thought process.

When faced with a decision, it might be beneficial to sacrifice a little of accuracy in cross-validation predictions in order to gain a stronger ability to explain when comparing two potential frameworks. In the instance of perovskite formability mentioned earlier, using a more intricate gradient increased DT tree resulted in a little decrease in the variability of the CV percentage points accuracy. However, this approach did not facilitate the creation of a straightforward visual representation. Utilizing data transformations like as Principal Component Analysis (PCA) to achieve higher levels of abstraction should be avoided when there is no explicit justification for the covariation of features in the input to models. A model that provides a clear rationale for the presence of uncertainty is less prone to unexpected outcomes in the future.

IV. DESIGNING A MATERIALS INFORMATICS MODEL

Due to the lack of a widely accepted procedure for addressing a materials informatics challenge, we endeavor to provide one potential approach, as has been done by Meredig [19]. The process may be categorized into four parts, as seen in **Fig 5**.

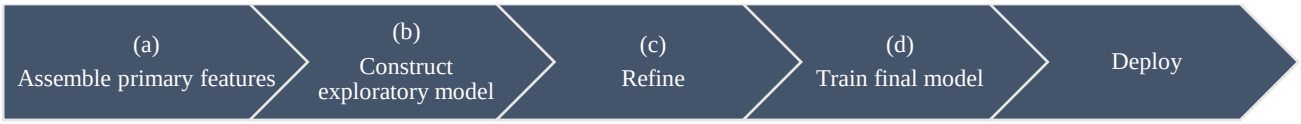

Fig 5. Diagrammatic representation of the sequential process involved in the design of a MI framework

The process begins with (A) the acquisition of information pertaining to the interest property. Subsequently, the process involves (B) constructing a basic model to investigate connections, followed by (C) enhancing the framework until it achieves a sufficient level of predicted accuracy, and (D) ultimately training and deploying the final model. Commencing with a simpler model facilitates deductive physical thinking rather than relying on explanations that come after the event.

To tackle any materials science challenge, it is essential to start by evaluating the existing domain knowledge (see **Fig 5a**). Subsequently, this expertise should be used to systematically list an initial collection of characteristics that will be gathered and analyzed for incorporation into a model. Selecting the appropriate main characteristics is often the most significant factor in determining the ultimate performance of the predictor. The primary function of machine learning algorithms is to identify the fundamental patterns within the data, rather than creating connections where there are none. The selection of major characteristics has many similarities with those of descriptors, as previously mentioned.

After assembling and preparing the main features for parsing, it is customary to create an exploratory framework using techniques such as a DT or RR. The purpose is to modify the features and determine which ones are most significant in the framework conceptualization (see **Fig 5b**). Analyze the impact of including or excluding main characteristics on the CV error and relate this to your previous knowledge and experience. It is advisable to avoid immediately using the most intricate models, such as kernel RR and NN. Doing so would require much more effort from both the researcher and the computer. Instead, it is more efficient to eliminate irrelevant components early on using domain expertise. Occasionally, a simple model adhering to Occam's razor is sufficient to provide the appropriate degree of accuracy and explanatory capability.

If there is no significant connection between the key characteristics and the output, beyond what would be anticipated by chance, then suggests that either the sample is flawed or the causative feature is absent from the data. Typically, if the samples number is ten times more than the number of characteristics and there is no noticeable link, it is probable that something is being overlooked. If the model's performance above chance levels but still has room for improvement, it suggests the presence of an underlying data pattern that requires further refining (see **Fig 5c**). In cases when there is a scarcity of samples of a certain kind, it may be necessary to gather more data or use other strategies to address the issue of underrepresented data. Excluding insignificant characteristics from initial testing might decrease the amount of irrelevant information that a model has to handle.

If there is a strong correlation between the characteristics, using a data dimension reduction approach such as Principal Component Analysis (PCA), as previously mentioned, may effectively reduce duplicate information. If there is a suspicion that applying a functional transformation to the data could be beneficial, such as for determined properties from feature extraction techniques, constitutive relationships can be employed. The results can then be reduced either through algorithmically or intuition using methods like feature ranking or LASSO from an ensemble algorithm like increased gradient decision trees.

Furthermore, the selection of the machine learning algorithm might also have an influence on the overall performance. An effective process would include many models of various sorts, which will be evaluated using cross-validation instead of selecting a single model initially. Every algorithm has its own set of advantages and disadvantages. Random forests are renowned for their high accuracy and robustness against outliers, yet they exhibit a slower training time when dealing with big datasets. Naive Bayesian classifiers have rapid training times, however provide untrustworthy probability estimations. Prior understanding of the advantages and disadvantages of a certain algorithm is highly suggested prior to its use.

After selecting hyperparameters and framework via refinement, the penultimate phase before deployment is to train the general data collection in order to hyper data encapsulated in the ultimate framework (see **Fig 5d**). The preceding cross-validation phase should give an error margin that aligns with existing realm knowledge [20]. By progressively increasing the complexity in a step-by-step manner, one may minimize the need for extensive backtracking to discover a simpler framework, which would be necessary if the process was less predictable. After achieving satisfactory model performance, it is crucial to analyze its significant discoveries and convert them into recommendations. Additionally, additional data should be created to enhance the description even further.

V. CONCLUSIONS

Overfitting is a prevalent issue in machine learning, characterized by algorithms excessively tailoring themselves to the training data and then struggling to effectively apply their knowledge to new data. Heuristic measures such as greedy search and gradient descent have been used to overcome this problem, but their effectiveness is often less than conventional methods. The selection of objective functions in optimization is of utmost importance, and empirical evidence suggests basic algorithms can naturally include corrective steps. It is also useful to include penalties in the algorithm. Decision trees are often too similar, and changing classification criteria can lead to different conclusions. In content reporting, the process of choosing annotations and extracting features is important for accurate forecasting and for gaining detailed knowledge of the underlying processes. Principal Component Analysis (PCA) is an effective technique for reducing the number of dimensions

in the data. However, caution is needed as this does not always ensure a clear separation of the data sets. Regularization methods such as Least Absolute Shrinkage and Selection Operator (LASSO) and ridge regression can effectively reduce overfitting in linear regression models. The choice of machine learning model influences its logic, where simple models such as decision trees, Bayesian methods can be described quickly. The process of content informatics modeling involves data acquisition, analytical model building, model letting go forward, examples of latest training and deployment scoring by cross-validation Should be done. Understanding the limitations and challenges associated with machine learning algorithms is essential for success in content data analysis.

Data Availability

No data was used to support this study.

Conflicts of Interests

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

Funding

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

Competing Interests

There are no competing interests.

References

[1]. R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, "Machine learning in materials informatics: recent applications and prospects," Npj Computational Materials, vol. 3, no. 1, Dec. 2017, doi: 10.1038/s41524-017-0056-5.

[2]. I. H. Sarker, "Machine learning: algorithms, Real-World applications and research directions," SN Computer Science, vol. 2, no. 3, Mar. 2021, doi: 10.1007/s42979-021-00592-x.

[3]. S. L. Kukreja, J. Löfberg, and M. Brenner, "A LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR (LASSO) FOR NONLINEAR SYSTEM IDENTIFICATION," IFAC Proceedings Volumes, vol. 39, no. 1, pp. 814–819, Jan. 2006, doi: 10.3182/20060329-3-au-2901.00128.

[4]. R. Gautam, S. Vanga, F. Ariese, and S. Umapathy, "Review of multidimensional data processing approaches for Raman and infrared spectroscopy," EPJ Techniques and Instrumentation, vol. 2, no. 1, Jun. 2015, doi: 10.1140/epjti/s40485-015-0018-6.

[5]. Ö. F. Alçın, A. Şengür, S. Ghofrani, and M. C. İnce, "GA-SELM: Greedy algorithms for sparse extreme learning machine," Measurement, vol. 55, pp. 126–132, Sep. 2014, doi: 10.1016/j.measurement.2014.04.012.

[6]. C. Tennant, A. Carpenter, T. Powers, A. Shabalina, L. Vidyaratne, and K. M. Iftekharuddin, "Superconducting radio-frequency cavity fault classification using machine learning at Jefferson Laboratory," Physical Review Accelerators and Beams, vol. 23, no. 11, Nov. 2020, doi: 10.1103/physrevaccelbeams.23.114601.

[7]. V. Srikant and D. R. Clarke, "On the optical band gap of zinc oxide," Journal of Applied Physics, vol. 83, no. 10, pp. 5447–5451, May 1998, doi: 10.1063/1.367375.

[8]. В. И. Анисимов, J. Zaanen, and O. K. Andersen, "Band theory and Mott insulators: HubbardUinstead of StonerI," Physical Review, vol. 44, no. 3, pp. 943–954, Jul. 1991, doi: 10.1103/physrevb.44.943.

[9]. S. Adams, O. Moretzki, and E. Canadell, "Global instability index optimizations for the localization of mobile protons," Solid State Ionics, vol. 168, no. 3–4, pp. 281–290, Mar. 2004, doi: 10.1016/j.ssi.2003.04.002.

[10]. Y. Okada et al., "Quasiparticle interference on cubic perovskite oxide surfaces," Physical Review Letters, vol. 119, no. 8, Aug. 2017, doi: 10.1103/physrevlett.119.086801.

[11]. K. J. Hubbard and D. G. Schlom, "Thermodynamic stability of binary oxides in contact with silicon," Journal of Materials Research, vol. 11, no. 11, pp. 2757–2776, Nov. 1996, doi: 10.1557/jmr.1996.0350.

[12]. C. Moure and O. Peña, "Recent advances in perovskites: Processing and properties," Progress in Solid State Chemistry, vol. 43, no. 4, pp. 123–148, Dec. 2015, doi: 10.1016/j.progsolidstchem.2015.09.001.

[13]. T. W. Schultz and M. Cronin, "Essential and desirable characteristics of ecotoxicity quantitative structure–activity relationships," Environmental Toxicology and Chemistry, vol. 22, no. 3, pp. 599–607, Mar. 2003, doi: 10.1002/etc.5620220319.

[14]. T. G. Dietterich and R. S. Michalski, "A Comparative Review of Selected Methods for Learning from Examples," in Springer eBooks, 1983, pp. 41–81. doi: 10.1007/978-3-662-12405-5_3.

[15]. W. Gao and X. Li, "Application of multi-task sparse lasso feature extraction and support vector machine regression in the stellar atmospheric parameterization," Chinese Astronomy and Astrophysics, vol. 41, no. 3, pp. 331–346, Jul. 2017, doi: 10.1016/j.chinastron.2017.08.004.

[16]. C. G. Bampis, L. Zhi, I. Katsavounidis, and A. C. Bovik, "Recurrent and dynamic models for predicting streaming video quality of experience," IEEE Transactions on Image Processing, vol. 27, no. 7, pp. 3316–3331, Jul. 2018, doi: 10.1109/tip.2018.2815842.

[17]. S. Wold, K. H. Esbensen, and P. Geladi, "Principal component analysis," Chemometrics and Intelligent Laboratory Systems, vol. 2, no. 1–3, pp. 37–52, Aug. 1987, doi: 10.1016/0169-7439(87)80084-9.

[18]. Z. Yong-Li and Y. Yang, "Cross-validation for selecting a model selection procedure," Journal of Econometrics, vol. 187, no. 1, pp. 95–112, Jul. 2015, doi: 10.1016/j.jeconom.2015.02.006.

[19]. B. Meredig, "Industrial materials informatics: Analyzing large-scale data to solve applied problems in R&D, manufacturing, and supply chain," Current Opinion in Solid State & Materials Science, vol. 21, no. 3, pp. 159–166, Jun. 2017, doi: 10.1016/j.cossms.2017.01.003.

[20]. B. Tsoi, R. Goeree, J. Jegathisawaran, J. Tarride, G. Blackhouse, and D. O'Reilly, "Do different decision-analytic modeling approaches produce different results? A systematic review of cross-validation studies," Expert Review of Pharmacoeconomics & Outcomes Research, vol. 15, no. 3, pp. 451–463, Mar. 2015, doi: 10.1586/14737167.2015.1021336.