Data-driven insights from predictive analytics on heterogeneous experimental data of industrial magnetic materials

Zijiang Yang*, Tetsushi Watari†, Daisuke Ichigozaki†,
Kei Morohoshi†, Yoshinori Suga†, Wei-keng Liao*, Alok Choudhary*, and Ankit Agrawal*
*Department of Electrical and Computer Engineering, Northwestern University
†Toyota Motor Corporation, Japan
*{zyz293, wkliao, choudhar, ankitag}@eecs.northwestern.edu,

***Abstract*—Data-driven methods are becoming increasingly popular in the field of materials science. While most data-driven models are trained on simulation data as it is relatively easier to collect a large amount of data from physics-based simulations, there are many challenges in applying data-driven methods on experiments: 1) experimental data is usually not clean; and 2) it generally has a greater degree of heterogeneity. In this project, we have developed a data-driven methodology to address these challenges on an industrial magnet dataset, where the goal is to predict magnetic properties (forward models) at different stages of the experimental workflow. The data-driven methodology consists of data cleaning, data preprocessing, feature extraction, and model development using traditional machine learning and deep learning methods to accurately predict magnet properties. In particular, we have developed three different types of predictive models: 1) numerical model using only numerical data containing composition and processing information; 2) image model using image data representing structure information; and 3) combination model using both types of data together. In addition to predictive models, the analysis and comparison of results across the models provide several interesting data-driven insights. Such data-driven analytics has the potential to help guide future experiments and realize the inverse models, which could significantly reduce costs and accelerate the discovery of new magnets with superior properties. The proposed models are already deployed in Toyota Motor Corporation.**

***Index Terms*—Deep learning, Gradient boosting, Heterogeneous data, Industrial magnet properties prediction, Materials informatics**

I. Introduction

The field of materials science relies on experiment and physics-based simulation to understand the underlying characteristics of different materials systems and design alternative materials for desired properties [1]–[4]. However, these conventional methods are not efficient. More specifically, experimentation is generally a trial-and-error method which is very expensive in terms of time and cost. Physics-based simulation is more efficient than experiment, but the simulation needs to solve the complex governing field equations for each material sample. Thus, it could still take prohibitively long to do the simulation of a large amount of material samples. In order to accelerate the process of materials discovery, the need for material informatics is emphasized by Materials Genome Initiative [5]. One of the advantages of using data-driven methods in materials science is its efficiency. Though in some cases it might take a long time to train the data-driven model, it is just a one-time effort. After the data-driven model is well trained, it can make accurate predictions in an efficient manner. In fact, data-driven methods have become increasingly popular in the field of materials science.

Traditional machine learning methods have gained great success in the prediction of materials' properties and design of materials system, such as steel fatigue strength prediction [6], mining of localization linkages [7] and machine learning system for multiscale materials science problems [8]. In recent years, deep learning method has shown its superiority to traditional machine learning methods, and has become a technique of choice in materials research, such as mining on homogenization linkages [9], electron microscopy image segmentation [10] and microstructural materials design [11]. Currently, most of the data-drive methods are based on data generated from physics-based simulation, because simulation data is usually clean and relatively easy to collect compared to experiment data. However, a simulation is still a proxy for experiment, as it only estimates the experimental outcome. Thus, performing actual experiments is considered the most accurate and trustable way to characterize materials. Past experiments therefore contain rich hidden information that needs to be uncovered and understood for get actionable insights, e.g., informing future experiments. Thus, how to effectively apply data-driven methods on experimental data has become an important research topic.

However, there are three main challenges to apply data-driven methods on experimental data: 1) Experimental data is usually not clean: Firstly, experimental data is generally quite noisy, because the measurement error could be introduced in different phases of experiment, such as error from machine or operations of researchers. Secondly, missing values are also common in experimental data. Thirdly, experimental data might contain outliers, which can be caused due to misoperation or incorrect settings of the machine; 2) Experimental data can be highly heterogeneous: Heterogeneity means consisting of different types of data. Especially in the materials science

2375-9259/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDMW.2019.00119

field, experimental data could have both numerical and image data, where numerical data records the information about composition of the material samples and processing parameters, and image data (e.g. Scanning Electron Microscope (SEM) image) represents the structural information of the material sample. 3) Experimental data is usually wide-shallow data: Wide means that experimental data usually contains many features, and shallow means the size of experimental data is relatively small. Thus, it is crucial to process such wide-shallow data in a strategic way to avoid the curse of dimensionality.

The above challenges make it more difficult to clean, extract salient information, and develop data-driven models based on experimental data. In this work, we have developed a data-driven framework to address the above challenges and develop accurate prediction models for magnet properties prediction. More specifically, in this work we focus on predicting four magnet properties separately based on the composition and processing information of magnet samples as well as SEM images indicating structural information of magnet samples. In other words, a strategic way for data cleaning, data preprocessing is introduced in this framework to process the noisy experimental data. A traditional machine learning and deep learning methods based method is proposed in the framework to train predictive models on the heterogeneous wide-shallow experimental dataset. In particular, three types of models are proposed, which are numerical model (i.e. purely using numerical data as input), image model (i.e. purely using image data as input) and combination model (i.e. using both numerical data and image data as input). The results show that the proposed framework can accurately predict magnet properties and some insights obtained from data analysis can help carry out the experiments in an efficient and effective manner, which might help to accelerate materials discovery. To the best of our knowledge, this is the first machine learning work on such heterogeneous industrial magnet data. In addition, the proposed framework can be easily extended to other materials systems which involve heterogeneous wide-shallow datasets, and it is already deployed in Toyota Motor Corporation.

II. Background and Machine Learning Methods

A. Magnet properties prediction

Due to global energy shortage and climate change, research on green energy has become a hot topic in recent decades. As renewable energy, electricity is widely used in different industrial applications. Thus, it is of significant importance to design better battery as well as its peripheral equipment, such that it has large capacity and efficient electric conversion rate with low cost. In this work, we have developed data-driven models for property prediction of magnetic materials that are located in the motors, sensors, and so on.

Coercivity (Hcj) and Remanence (Br) are the two main metrics of interest to measure the performance of a magnet sample. In order to enhance Hcj and Br, some processing steps are applied on the magnet samples as shown in Figure 1. More specifically, the raw magnet samples undergo four different processing steps, which are rapid cooling, molding, hot deform and heat treatment. During the processing, the properties (i.e. Hcj and Br) are measured twice separately after hot deform and heat treatment processing steps, and SEM images are taken after hot deform (Image) processing step to visualize the structural information of magnet samples. Thus, given composition and processing information as well as structural information, prediction models are developed in this work to accurately predict two sets of magnetic properties (i.e. Hcj and Br of Property No.1 and No.2 as shown in Figure 1).

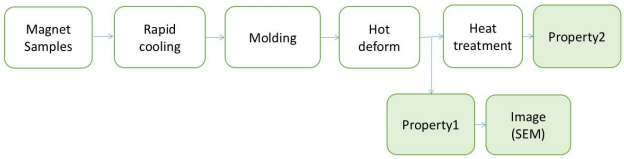

Fig. 1. Demonstration of processing steps of magnet samples.

B. Gradient boosting

Boosting is one of the most commonly used machine learning methods, and it has been widely used in various research tasks [12]–[14]. Particularly, gradient boosting [15] is a variant of boosting method. Gradient boosting involves three elements: 1) Loss function: One of the advantages of gradient boosting is that it is a generic framework, which means it can use a variety of loss functions. In particular, squared error is a commonly used loss function for regression problems. 2) Weak learner: Decision tree is typically used as the weak learner in gradient boosting. Decision tree is constructed in a greedy manner, which chooses the features that can minimize the loss after splitting the node. 3) Additive model: Gradient boosting is an iterative training method, which means it adds a decision tree to the model to reduce the loss in each training iteration. The training of gradient boosting is stopped when a predefined number of decision trees are added to the model, or the loss reaches an acceptable level. In this work, gradient boosting regressor, which is a gradient boosting method for regression problems, is used to train the models for magnetic properties prediction.

C. Transfer learning

In order to train a successful deep learning model, a large amount of data is usually required. However, in some scientific domains, such as materials science, it is very expensive to collect such large amount of labeled data, which hinders the application of deep learning. To take advantage of deep learning even with a relatively small dataset, one of the common approaches is to use transfer learning [16]. Transfer learning focuses on applying the knowledge learned from solving one problem to another different but related problem. More specifically, using pre-trained models as a feature extractor is one of the most common approaches for transfer learning. Because of the hierarchical learning strategy of convolutional

neural network, it can detect simple features, such as edges, at earlier layers, then later layers combine them to form some high level features. Since the pre-trained model is trained on a huge amount of various types of images, the learned features have the ability to well characterize different types of images. Thus, the pre-trained model can be used as a feature extractor to extract features from images so that machine learning methods could be further applied on the extracted features to train the prediction model.

In this work, we use a portion of VGG-16 [17] network as a feature extractor. As shown in Figure 2, VGG-16 consists of five convolutional blocks and two fully connected layers. As mentioned above, the convolutional blocks (i.e. blue blocks in Figure 2) are used to learn the features that can well characterize images, while the fully connected layers (i.e. red block in Figure 2) use the learned features to make the prediction. In this work, we keep the architecture and weights of all the convolutional blocks of VGG-16 as feature extractor to extract features from SEM images of magnet samples. Then gradient boosting regressor is trained on the extracted features to make the predictions.

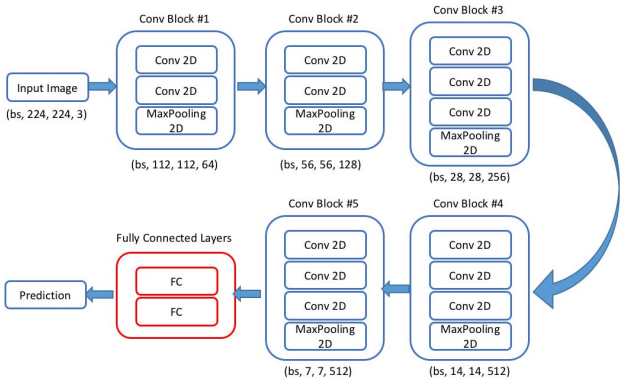

Fig. 2. Architecture of VGG16 pre-trained model. (bs. is the abbreviation of batch size)

III. Dataset

This dataset is collected from experiments by Toyota Motor Corporation, Japan for magnet properties research, and it includes numerical data and image data.

A. Numerical data

As shown in Figure 1, there are four processing steps (i.e. rapid cooling, molding, hot deform and heat treatment) to improve the mechanical properties of a magnet sample. Thus, the numerical data mainly contains the processing information of four processing steps, such as processing time, processing temperature and pressure. The composition of the magnet sample is also included in the numerical data. In addition, at most four magnet properties can be measured for a magnet sample. More specifically, Property No.1 (i.e. P1 Hcj and P1 Br) are measured after hot deform processing step, while Property No.2 (i.e. P2 Hcj and P2 Br) are measured after heat treatment processing step.

B. Image data

Images of the magnet samples are taken after hot deform processing step using SEM technique. For each magnet sample, SEM images are taken at up to eight positions with three different magnifications (i.e. $\times 200$, $\times 1000$, $\times 30000$) and two different SEM modes (i.e. COMPO and SEI). In other words, there are at most 48 ($= 8 \times 3 \times 2$) SEM images for a magnet sample. Figure 3 (a) shows the two most common positions where SEM images are available in the dataset, the position $C00$ is the center position of the magnet sample, while position $C10$ is the position shifted along Z axis compared to position $C00$. Figure 3 (b) shows an example of SEM image taken at $C00$ position with $\times 1000$ magnification and COMPO SEM mode.

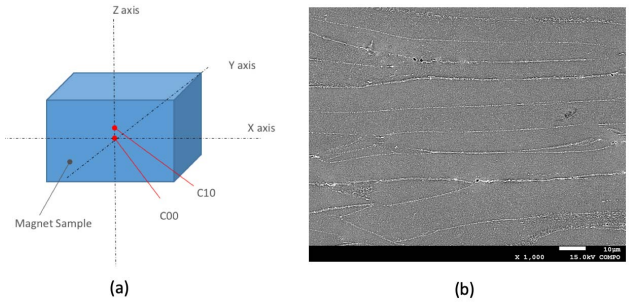

Fig. 3. (a) Illustration of the positions where SEM images are taken at a magnet sample. (b) An example of SEM image taken at $C00$ position with $\times 1000$ magnification and COMPO SEM mode.

IV. Methods

A. Data preprocessing

Because the dataset is collected from experiments, there are noise, missing values and outliers in the dataset. In order to train an accurate prediction model, data preprocessing is necessary.

1) Data preprocessing for numerical data: For numerical data, the data preprocessing, such as feature removal and outlier detection, is mainly based on domain knowledge. Particularly, a four-step preprocessing method is applied on numerical data to obtain the corresponding dataset.

- Numerical data includes the features representing composition and processing information of magnet samples. Since the dataset is relatively small, such large number of features might lead to the curse of dimensionality. Thus, we remove features that are not related to magnet properties (i.e. features that do not effect properties based on domain knowledge) and features that are correlated with others (i.e. features that can be calculated based on other features). For example, manufacturing device name is removed because it is not related to magnet properties, and the aspect ratio of a sample is removed since it is correlated with its width and length.
- Then we remove the data points without corresponding magnet property.

- In the experiments, the property of a magnet sample is measured multiple times, and the magnet property of the same magnet sample might be slightly different due to the measurement errors of the machine. Thus, we remove the outliers (i.e., where the magnetic property is significantly different from other same magnet sample), fill up the missing values with the average value from the same magnet sample, and take the average value of the magnet property of the same magnet sample as the value of its property.
- The range of different features are highly diverse, so we rescale each feature individually to unit norm.

After data preprocessing, there are 43 features left and they can be categorized into four categories:

- Composition: The magnet sample consists of 9 chemical elements.
- Dimension: There are 4 features describing the physical dimension of the magnet sample before hot deform processing step.
- Phase: There are 18 features representing the information of different phases in the rapid cooling processing step.
- Processing parameters: There are 12 features in total representing the processing parameters of the rest three processing steps. More specifically, molding, hot deform and heat treatment processing steps include 4, 5 and 3 features, respectively.

Note that after data preprocessing, all the 43 features are available for predicting Property No.2, while only 40 features are available for predicting Property No.1, since Property No.1 is measured before heat treatment processing step is applied on the magnet sample.

2) Data preprocessing for image data: As shown in Figure 3 (b), the SEM image has label information at the bottom, which does not contain any structure information of the magnet sample. Thus, the bottom of each SEM image is cropped. Since the sizes of SEM images of different magnifications are different, we then resize the SEM images to 224×224. Finally, the number of data points for each magnet property after data preprocessing is listed in Table I.

TABLE I
THE NUMBER OF DATA POINTS FOR EACH MAGNET PROPERTY AFTER DATA PREPROCESSING

Property	Number of data points
Property No.1 - Hcj	98
Property No.1 - Br	98
Property No.2 - Hcj	107
Property No.2 - Br	99

B. Proposed models

We compare the performance of different combinations of machine learning methods such as gradient boosting, random forest and support vector machine as regressors and pre-trained networks such as VGG-16, VGG-19 [17] and ResNet [18] as feature extractors. For each regressor, we performed an extensive grid search for optimization of hyperparameters to find the best hyperparameters. For instance, for gradient boosting regressor, we used a learning rate from [0.01, 0.1, 0.5], number of estimators from [50,100,150,200] and maximum depth from [1, 3, 5, 10]. Similarly, for each pre-trained networks, different layers are tried as feature extractor to find the best image representative for current application. Among all of the combinations of regressors and feature extractors, the proposed models that use gradient boosting regressor and VGG-16 as feature extractor perform the best. In the next sections, three types of models based on gradient boosting regressor and VGG-16 are proposed for each magnet property, which are numerical model, image model and combination model.

1) Numerical model: Numerical model (referred as Num model) takes numerical data as input and uses gradient boosting regressor to train the prediction model. More specifically, gradient boosting regressor with 0.1 learning rate, 100 estimators and maximum depth of 3 is used. The same parameter settings of gradient boosting regressor is used for both image model and combination model, which are introduced later. As mentioned in section 3, the number of features for different magnet properties are different. More specifically, the number of input features of numerical model for Property No.1 and No.2 are 40 and 43, respectively.

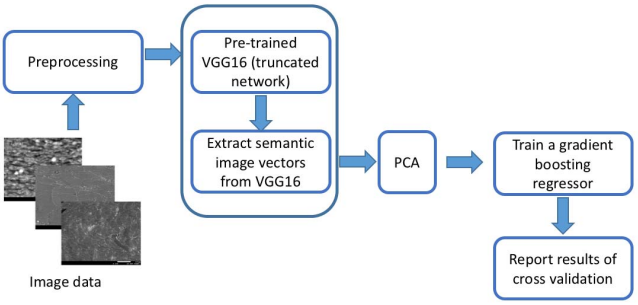

Fig. 4. The flowchart of the proposed image models.

2) Image model: After analyzing the image data, we find that using either three images (i.e. images of position $C00$ with 3 magnifications and COMPO SEM mode) or six images (i.e. images of position $C00$ and $C10$ with 3 magnifications and COMPO SEM mode) gives the best performance. Thus, two image models are proposed in this work. One image model (referred as 3M model) takes three SEM images of a magnet sample as input (i.e. images of position $C00$ with 3 magnifications and COMPO SEM mode), while the other image model (referred as 6M model) takes six SEM images of a magnet sample as input (i.e. images of position $C00$ and $C10$ with three magnifications and COMPO SEM mode, respectively). The flowchart of both image models are the same, and it is shown in Figure 4, and it includes three steps:

- After image preprocessing, we use transfer learning technique to extract semantic image vectors from images. More specifically, we truncate VGG-16 by keeping all the

convolutional blocks (i.e. blue blocks in Figure 2), and feed the preprocessed image into the truncated network to get the output (i.e. 512 features maps) of the last convolutional block.

- However, the dimensionality of the output of the last convolutional block is too high, which might lead to the curse of dimensionality. Thus, we apply two methods to reduce the dimensionality. First, global average pooling [19] is applied on each feature maps individual by computing the average of entries' values of each feature map. In this way, we could convert these 512 feature maps to a 1-D vector with 512 entries and it is the representation of the input image. Since the image model takes multiple SEM images as input, we compute such vector for each input image individually and concatenate them together into one 1-D vector. However, the dimensionality of this 1-D vector is still high compared to the number of features of numerical data, so Principal Component Analysis (PCA) [20] is applied to further reduce the dimension to 25. In particular, the summation of explained variance ratio of the selected principal components are around 91% and 87% for 3M and 6M model, which implies the selected principal components contain the enough information of images. In other words, by using transfer learning and dimensionality reduction techniques, we could use a 1-D semantic image vector with 25 entries to represent input images.
- Finally, gradient boosting regressor takes the semantic image vectors as input to train the prediction model for each magnet property.

3) Combination model: The combination model takes both numerical data and image data as input, and two combination models are proposed in this work. The difference of the two combination models is the number of input SEM images. One combination model (referred as 3NM model) takes three SEM images of a magnet sample (i.e. images of position $C00$ with 3 magnifications and COMPO SEM mode) and numerical data as input, while the other combination model (referred as 6NM model) takes six SEM images of a magnet sample (i.e. images of position $C00$ and $C10$ with three magnifications and COMPO SEM mode, respectively) and numerical data as input. The flowchart of combination model is shown in Figure 5. More specifically, the method to compute the semantic image vector from input images is the same as the image model. Then the semantic image vector is concatenated with numerical data, and they are fit to each magnet property using gradient boosting regressor.

V. Results and Discussion

A. Experimental setting and error metric

We use two error metrics to evaluate the performance of the proposed models, which are mean absolute error rate (MAE%)

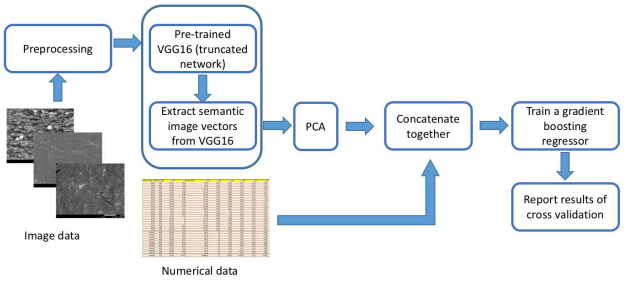

Fig. 5. The flowchart of the proposed combination models.

and pearson correlation coefficient (R). The equations for computing three metrics are shown as below,

$$MAE\% = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i} \times 100 \quad (1)$$

$$R = \frac{\sum_{i=1}^{N}(y_i - m_y)(\hat{y}_i - m_{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - m_y)^2(\hat{y}_i - m_{\hat{y}})^2}} \quad (2)$$

where N is the total number of data points in the dataset of corresponding magnet property. y_i and $\hat{y}_i$ represent the ground truth and predicted values, respectively. m_y and $m_{\hat{y}}$ denote the mean of the ground truth and predicted values, respectively. In addition, because the dataset is relatively small, 5-fold cross validation is implemented to evaluate performance of all the proposed models.

B. Results analysis and data-driven insights

In this section, the performance of the proposed models are evaluated and data-driven insights are discussed based on the experimental results. The proposed models are trained for each magnet property, and the results of 5-fold cross validation in terms of MAE% and R are shown in Table II and Table III. From the results, we can observe that numerical model can already achieve a very high accuracy although it only uses numerical data as input. More specifically, the numerical model can get 3.56% and 2.20% MAE% for Property No.1 Hcj and Br, respectively. For Property No.2 Hcj and Br, the numerical model can achieve 4.23% and 2.98% MAE%. Meanwhile, the results of Table III follows the same trend.

TABLE II
PERFORMANCE COMPARISON OF THE PROPOSED MODELS IN TERMS OF MAE%

Property	Num model	3M model	6M model	3NM model	6NM model
P1 Hcj	3.56%	9.56%	7.71%	4.07%	3.97%
P1 Br	2.20%	2.36%	2.35%	2.12%	2.07%
P2 Hcj	4.23%	13.00%	12.46%	6.00%	5.69%
P2 Br	2.98%	3.35%	3.32%	2.53%	2.48%

Figure 6 shows the parity plot of all the four magnet properties based on numerical model. The plots show that most of data points are distributed along the diagonal, which shows

TABLE III
PERFORMANCE COMPARISON OF THE PROPOSED MODELS IN TERMS OF R

Property	Num model	3M model	6M model	3NM model	6NM model
P1 Hcj	0.93	0.65	0.78	0.93	0.93
P1 Br	0.60	0.56	0.57	0.63	0.64
P2 Hcj	0.98	0.78	0.81	0.96	0.96
P2 Br	0.45	0.35	0.39	0.55	0.57

that the proposed model can make accurate predictions. However, the variance of parity plot of Property Br is larger than that of Property Hcj. In addition, it is relatively straightforward to retrieve importance scores for each feature after gradient boosting regressor is trained. In particular, feature importance is calculated for a weak learner by the amount that each feature split point improves the performance measure. The feature importance is then averaged across all of the weak learners within the model. Figure 7 shows the feature importance for each magnet property based on the numerical model.

- **Insights (1)**: The four processing steps have different purposes. Rapid cooling is to make crystallization and control the size of crystallization. Molding is to increase sample's density. Hot deform aligns the orientation of the crystal to get high property. Heat treatment smooths the unevenness of grain boundaries of sample's structure to get high property. Thus the hot deform and heat treatment processing steps are the crucial processing to enhance the mechanical property of the magnet samples. In Figure 7, the features of hot deform and heat treatment processing steps are found to be more important than other features, which matches the domain knowledge.

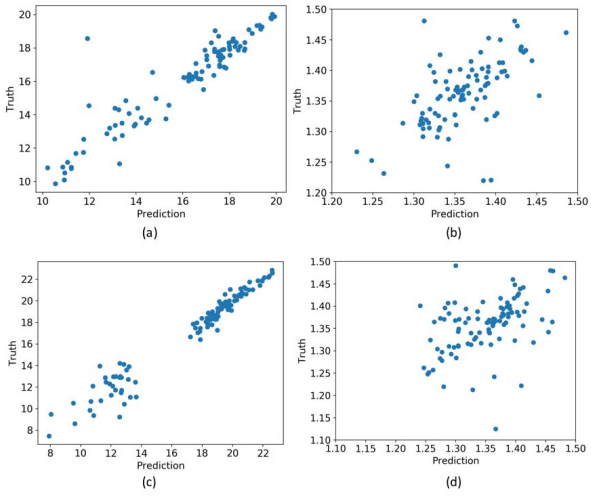

Fig. 6. The parity plots of each magnet property based on numerical model. (a) Property No.1 Hcj. (b) Property No.1 Br. (c) Property No.2 Hcj. (d) Property No.2 Br

However, the performance of both image models is worse than the corresponding numerical models, and the performance of 6M model for all the magnet properties is slightly better than that of 3M model. The reasons might be twofold: 1) The SEM technique would destroy the sample after taking the image. So although the SEM image and magnet property could be collected from the samples with same composition and processing parameters, there is no one-to-one mapping between SEM image and magnet property. In other words, the SEM images and corresponding magnet property are not measured from exactly the same sample, which could introduce noise in the dataset. 2) There are six images (i.e. positions $C00$, $C10$ with three magnifications and COMPO SEM mode) available for each magnet sample. By taking more SEM images as input, the model can learn more knowledge about structural information. Thus, the performance of 6M models are better than that of 3M models. Another three insights could be found be analyzing and reasoning the results of the proposed image models:

- **Insights (2)**: From data mining point of view, it is important to retain a one-to-one mapping between model's input and output. Since SEM techniques could destroy the samples, it is important to change the experimant operations order that the properties of samples should be measured before SEM images are analyzed.
- **Insights (3)**: As shown in Figure 3 (a), the position $C00$ is the center position of the magnet sample, while position $C10$ is the position shifted along Z axis compared with position $C00$. Since both positions are at the center area of the magnet sample, the SEM images of the two positions might contain similar structural information. Thus, only slight performance improvement can be observed when we include more SEM images from position $C10$. The performance might be improved if more SEM images from other different sample areas are available.
- **Insights (4)**: In addition, we can observe that the performance of the image models for Property Hcj are worse than that of image models for Property Br. Interestingly, this might indicate that the structural information in the SEM images is more related to the property Br, but we are not aware of any existing domain knowledge supporting or refuting this data-driven insight.

The performance of 6NM model for all the magnet properties is slightly better that that of 3NM model, and the reasons for the improvement are the same as mentioned above. In addition, the performance of the combination model has different trends for Property Hcj and Br when compared with that of the numerical model. As mentioned in insights (4), it appears that the structural information in the SEM images is more informative for Property Br, so including SEM images in the combination model leads to a performance improvement compared to the numerical models for Property Br. On the other hand, since the performance of the image model is much worse than that of numerical model for Property Hcj, adding SEM images in the combination model does not improve the performance, but rather deteriorates the performance as it might confound the model.

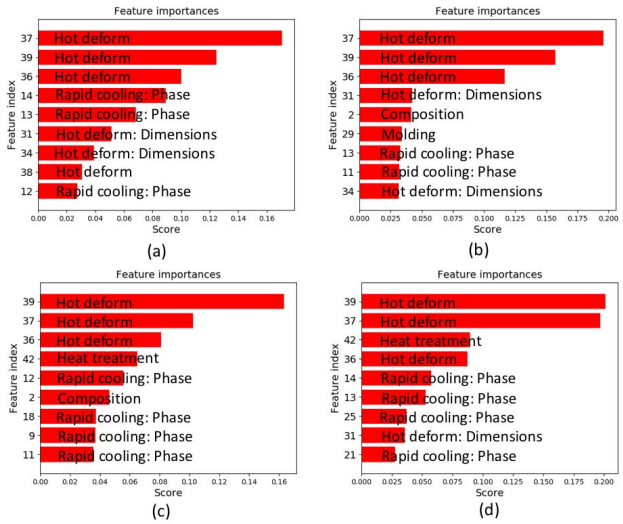

Fig. 7. The features importance for each magnet property based on numerical model. (a) Property No.1 Hcj. (b) Property No.1 Br. (c) Property No.2 Hcj. (d) Property No.2 Br.

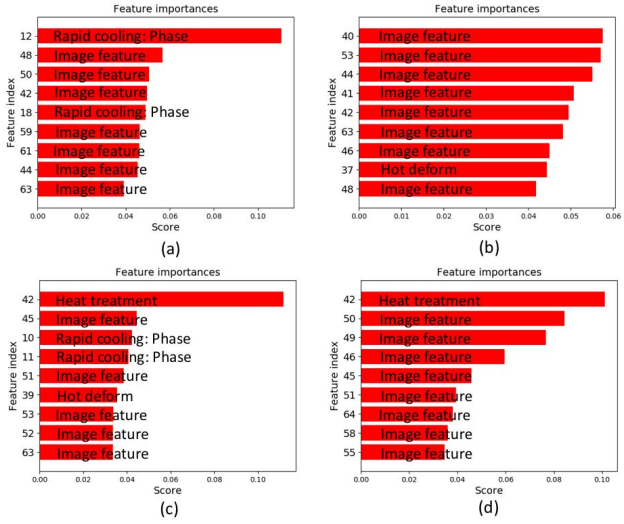

Fig. 8. The features importance for each magnet property based on 3NM model. (a) Property No.1 Hcj. (b) Property No.1 Br. (c) Property No.2 Hcj. (d) Property No.2 Br

Figures 8 and 9 show the feature importance for each magnet property based on 3NM and 6NM model, respectively. By comparing with Figure 7, we can observe that several image features turn out to be important to predict the magnetic properties. However, the importance of image features is different for Property Hcj and Br.

- **Insights (5)**: We can observe that features of numerical data are more important than features of image data when predicting Property Hcj, while image features are much more important than numerical features for Property Br. Thus, different types of data and models should be used depending on the property of interest.

This finding also supports our earlier mentioned data-driven insight (4) that the SEM image contains more information related to Property Br than Property Hcj, and that might also be the reason why the image and combination models performs better for Property Br than Property Hcj.

VI. Applications for different materials systems and predictive experimental design

Experiments data, especially industrial experiments data, is significantly different from simulation data, because it is usually a noisy heterogeneous wide-shallow dataset. Thus, a strategic way to process such dataset is important in order to develop an accurate machine learning model. The proposed framework in this work is general enough to be extend to other industrial experiments dataset. More specifically, different data resources should be processed in different ways in data cleaning and data preprocessing. Numerical data is extremely noisy, since it records all the information during the manufacturing process. Thus, domain knowledge is required to process numerical data, such as feature removal and outlier detection. On the other hand, the goal of data preprocessing for image data is to avoid the curse of dimensionality due to the limited

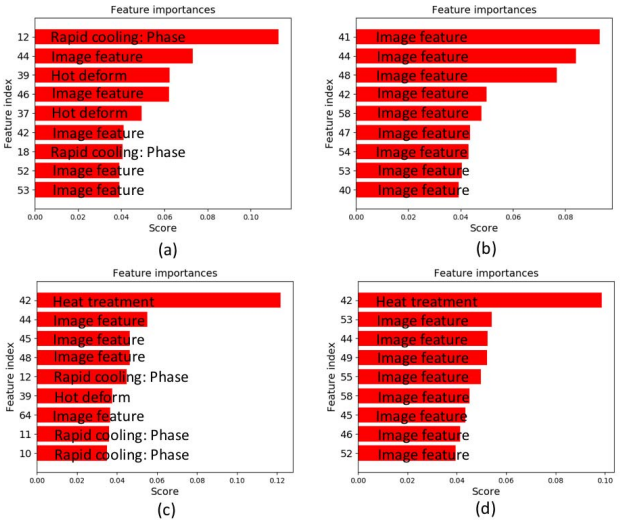

Fig. 9. The features importance for each magnet property based on 6NM model. (a) Property No.1 Hcj. (b) Property No.1 Br. (c) Property No.2 Hcj. (d) Property No.2 Br

experiments data, and it is mainly based on machine learning methods. Due to the different characteristics of materials systems, different pre-trained models and dimension reduction techniques can be replaced in the framework to extract salient features from image data. The choice of machine learning models in model development is highly dependent on the goal of problem. Traditional machine learning method, such as gradient boosting, is easier to interpret so that more insights might be discovered, while deep learning models usually has higher learning capability if a relatively large dataset is available for training.

The proposed methods are already deployed in Toyota

Motor Corporation, and the benefits are twofold. As mentioned above, data-driven insights have been obtained from analyzing the results. In general, a successful data mining project should result in several data-driven insights, most of which should be in agreement with domain knowledge and intuition (things we already know), but it should also include some surprises (things that we do not know). The reconfirmation of known knowledge gives us overall confidence in the correctness of the data mining model, whereas surprises provide valuable hints towards new not-yet-known knowledge, which can open up new avenues for further investigation and actually lead to knowledge discovery. We believe insights obtained from the results include a healthy mix of known and not-yet-known knowledge.

In the long term, such materials informatics approaches have the potential to significantly reduce costs by guiding future experiments, as well as accelerate the discovery of new magnets. Experiments are extremely expensive in terms of time, manpower, and money. In addition, experiment is a trial-and-error process that is usually conducted mainly based on operator's experience so that it could be inefficient. Machine learning guided experimentation can help avoid both unnecessary experiments as well as high-end processing/characterization of not-so-promising magnet samples/locations, subsequently also reducing the man-hours required to perform experimentation and operation of SEM imaging equipment. Moreover, it can also help in narrowing down the infinite search space of possible magnets by prescreening and subsequent exploration of the most promising regions. On the other hand, the new data obtained from experiments guided by machine learning can be used to refine the predictive model to improve prediction accuracy. Thus, such materials informatics approaches provides unprecedented opportunities for significantly accelerating the discovery and design of new magnets (or more broadly materials) with superior properties.

VII. Conclusion

In this work, we present a data-driven methodology to develop prediction models for magnet properties on experimental dataset. The results show that the numerical model can provide accurate predictions, and adding SEM image in the model could improve the model's performance when important structural information is contained in the image. There are several future directions for each type of prediction model. 1) Numerical model: By designing advanced features based on the raw features, the performance might be improved. 2) Image model: By applying aggressive data augmentation method, deep learning method might provide better prediction. 3) Combination model: Fine-tuning on the pre-trained model might be a promising method to extract problem-specific semantic image vector for microstructure, which could be combined in different ways with the numeric data in order to build a more accurate prediction model.

Acknowledgement

This work was supported by Toyota Motor Corporation.

References

[1] A. Agrawal and A. Choudhary, "Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science," *APL Materials*, vol. 4, no. 053208, pp. 1–10, 2016.

[2] S. Nguyen, A. Tran-Le, M. Vu, Q. To, O. Douzane, and T. Langlet, "Modeling thermal conductivity of hemp insulation material: A multi-scale homogenization approach," *Building and Environment*, vol. 107, pp. 127–134, 2016.

[3] X.-Y. Zhou, P. Gosling, C. Pearce, Z. Ullah *et al.*, "Perturbation-based stochastic multi-scale computational homogenization method for the determination of the effective properties of composite materials with random properties," *Computer Methods in Applied Mechanics and Engineering*, vol. 300, pp. 84–105, 2016.

[4] A. Cruzado, B. Gan, M. Jiménez, D. Barba, K. Ostolaza, A. Linaza, J. Molina-Aldareguia, J. Llorca, and J. Segurado, "Multiscale modeling of the mechanical behavior of in718 superalloy based on micropillar compression and computational homogenization," *Acta Materialia*, vol. 98, pp. 242–253, 2015.

[5] N. Science and T. C. (US), *Materials genome initiative for global competitiveness*. Executive Office of the President, National Science and Technology Council, 2011.

[6] A. Agrawal, P. D. Deshpande, A. Cecen, G. P. Basavarsu, A. N. Choudhary, and S. R. Kalidindi, "Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters," *Integrating Materials and Manufacturing Innovation*, vol. 3, no. 8, pp. 1–19, 2014.

[7] R. Liu, Y. C. Yabansu, Z. Yang, A. N. Choudhary, S. R. Kalidindi, and A. Agrawal, "Context aware machine learning approaches for modeling elastic localization in three-dimensional composite microstructures," *Integrating Materials and Manufacturing Innovation*, pp. 1–12, 2017.

[8] D. Wheeler, D. Brough, T. Fast, S. Kalidindi, and A. Reid, "PyMKS: Materials Knowledge System in Python," 5 2014. [Online]. Available: https://figshare.com/articles/pymks/1015761

[9] Z. Yang, Y. C. Yabansu, R. Al-Bahrani, W.-k. Liao, A. N. Choudhary, S. R. Kalidindi, and A. Agrawal, "Deep learning approaches for mining structure-property linkages in high contrast composites from simulation datasets," *Computational Materials Science*, vol. 151, pp. 278–287, 2018.

[10] M. L. Silvoster and V. Govindan, "Enhanced cnn based electron microscopy image segmentation," *Cybernetics and Information Technologies*, vol. 12, no. 2, pp. 84–97, 2012.

[11] Z. Yang, X. Li, L. C. Brinson, A. N. Choudhary, W. Chen, and A. Agrawal, "Microstructural materials design via deep adversarial learning methodology," *Journal of Mechanical Design*, vol. 140, no. 11, p. 111416, 2018.

[12] H. Deng, S. P. Eckel, R. Urman, K. Berhane, and F. Gilliland, "Predicting bronchitic symptoms using gradient boosting models for longitudinal data," in *B106. PRENATAL, PERINATAL, AND CHILDHOOD EXPOSURES IN LUNG DISEASE*. American Thoracic Society, 2017, pp. A4805–A4805.

[13] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 308–324, 2015.

[14] R. P. Sheridan, W. M. Wang, A. Liaw, J. Ma, and E. M. Gifford, "Extreme gradient boosting as a method for quantitative structure–activity relationships," *Journal of chemical information and modeling*, vol. 56, no. 12, pp. 2353–2360, 2016.

[15] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[16] K. Gopalakrishnan, S. K. Khaitan, A. Choudhary, and A. Agrawal, "Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection," *Construction and Building Materials*, vol. 157, pp. 322–330, 2017.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[20] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.