Estimation of Mechanical Properties of Mancos Shale using Machine Learning Methods

Teeratorn Kadeethum and Hongkyu Yoon

Geomechanics Department, Sandia National Laboratories, Albuquerque, NM, 87185, USA

ABSTRACT: We propose the use of balanced iterative reducing and clustering using hierarchies (BIRCH) combined with linear regression to predict the reduced Young's modulus and hardness of highly heterogeneous materials from a set of nanoindentation experiments. We first use BIRCH to cluster the dataset according to its mineral compositions, which are derived from the spectral matching of energy-dispersive spectroscopy data through the modular automated processing system (MAPS) platform. We observe that grouping our dataset into five clusters yields the best accuracy as well as a reasonable representation of mineralogy in each cluster. Subsequently, we test four types of regression models, namely linear regression, support vector regression, Gaussian process regression, and extreme gradient boosting regression. The linear regression and Gaussian process regression provide the most accurate prediction, and the proposed framework yields R^2 = 0.93 for the test set. Although the study is needed more comprehensively, our results shows that machine learning methods such as linear regression or Gaussian process regression can be used to accurately estimate mechanical properties with a proper number of grouping based on compositional data.

1 INTRODUCTION

The hydro, mechanical, and chemical properties of shale formations with compositional and textural heterogeneity across a range of scales give rise to very complex behavior under various environmental and engineered conditions. Various geologic variables, including mineralogy, types of cement, organic content, and the spatial distribution of these characteristics, contribute to mechanical properties (elastic properties, fracture toughness, anisotropy, etc.). These compositional and structural heterogeneity in very-fine sedimentary rocks may affect the onset and propagation of brittle fracture in shale and can lead to the formations of flow conduit. Given the formation and operational conditions (e.g., stress, natural fractures, injection fluid and pressure) the geometry and extent of fracture networks is predominantly determined by shale mechanical properties.

In our previous work (Yoon et al., 2020) on multiscale mechanical properties of Mancos shale, Young's moduli at microscale based on nanoindentation were larger than those at the laboratory core scale, which is caused by the combined effect of composition, textures, and interfaces of mineral phases. Analysis of mineralogy distribution based on MAPS (Modular Automated Processing System) technique and detailed petrographic analysis of tested samples reveal the important effect of compositional distribution, micropores, and bedding boundaries on the patterns of microfracture propagation. However, estimated Young's modulus values using composition-based mixture models do not match the measured values. Instead, measured Young's modulus and hardness values have much higher correlations when nanoindentation data are grouped into three categories such as strong minerals (quartz, feldspar, and pyrite), carbonates (calcite, dolomite), and clay-rich group. With high resolution scanning electron imaging of indentation markers these data sets show how the mechanical response during indentation is influenced by compositional distribution at the indentation location.

In this work We employ multiple machine learning methods to estimate Young's modulus of Mancos shale using nanoindentation experimental data. Detailed mineralogy data from the MAPS mineralogy provides accurate compositional data over the indentation area. This work allows us to develop robust predictive model of estimating

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

mechanical properties based on compositional data.

2 METHODOLOGY

2.1. Experimental methods

A Mancos shale sample selected for this work is a Cretaceous shale located in the western United States/Rocky Mountains. Detailed mechanical properties and compositional analysis work from a large quarry Mancos Shale block (12 inches high and 15 inches in diameter) was reported in the previous work (Yoon et al., 2020). The Mancos shale used in this work has a relatively low total organic carbon (less than 1-2 % by weight) and contains the fine scale of interbedding of muds and sands. For micron-scale characterization, 2 mm thick sample was prepared from a small cylindrical core sample for mechanical testing. The sample was polished by argon ion milling (Fischione 1060 SEM Mill) and then analyzed with backscattered electron scanning (BSE) and energy dispersive spectroscopy (EDS). MAPS (Modular Automated Processing System) Mineralogy platform was used for SEM-based automated mineralogical measurement, analysis, data integration. With spectral matching of EDS data each pixel can be identified as single or multiple minerals. In this work we employed BSE and EDS analysis at 0.2 and 2 micron resolution, respectively.

Nanoindentation was conducted using a Hysitron TriboIndenter 900 with a Berkovich geometry diamond tip over multiple regions selected based on BSE and EDS mapping in 2cm and 1.5 cm area. A 5 x 5 grid array of standard indents spaced 20 um apart with an indentation strain rate of 0.1 (Lucas et al., 1996) and a maximum load of 10 mN. Hardness and Young's modulus measurement over a total of nine different arrays were performed to provide 225 data points. Hardness was computed by the maximum load over the contact area and the reduced Young's modulus was computed with the stiffness calculated as the slope of initial unloading, a geometrical constant of the Berkovich tip, and contact area. After indentation testing, SEM image of each indentation impression was obtained using a FEI Helios Nanolab G3 CX DualBeam FIB/SEM. An example of these images is shown in Figure 1

The mineral composition used in this study is listed in Table 1. There are 17 types of mineral, the *unclassified* label to represent a part of SEM images that we cannot define, the *Porosity* label represents a void inside a porous media, and organics.

Table 1: List of mineral composition

mineral composition			
quartz	feldspar	muscovite	kaolinite
illite	smectite	Mg-chlorite	Fe-chlorite
zircon	calcite	dolomite	ankerite
apatite	monazite	pyrite	sphalerite
rutile	unclassified	porosity	organics

2.2. Machine learning methods

We use the balanced iterative reducing and clustering using hierarchies (BIRCH) to cluster our dataset through the mineral compositions shown in Table 1 (Zhang et al., 1996). BIRCH constructs a tree structure from which cluster centroids are extracted. To elaborate, it clusters incoming multi-dimensional metric data points to produce the best quality clustering. BIRCH has two primary hyperparameters, *threshold* and *number of clusters*. The threshold constraints a radius of the sub-cluster obtained by merging a new sample and the closest sub-cluster. We set it as 0.001 throughout this study. The number of clusters is self-explanatory, and in short, it represents the number of clusters after the final clustering step (i.e., the final number of clusters). We utilize the BIRCH implementation provided by Pedregosa et al. (2011).

Throughout this study, we test four types of regression models; 1. linear regression (LR), 2. support vector regression (SVR), 3. Gaussian process regression (GP), and 4. extreme gradient boosting regression (XGBoost). We will briefly summarize these models in the following paragraphs. We multiple LR in this study, which means it uses *many* explanatory, or independent, variables to predict the outcome of *one* response, or dependent, variable. We use an API provided by Buitinck et al. (2013).

For the SVR, which also takes many explanatory variables; and subsequently maps them to one response. We utilize a quadratic polynomial kernel with an independent term of one, a regularization parameter (C) of 100, and kernel coefficient (Γ) of one over number explanatory variables. Again, we employ an API provided by Buitinck et al. (2013). For the GP, the model itself is also a multiple regressor, which means it can take many independent variables and predict one dependent variable. We use an API developed by Buitinck et al. (2013) with Dot-Product kernel given by

$$k\left(x_i, x_j\right) = \sigma_0^2 + x_i \cdot x_j, \tag{1}$$

where σ_0^2 is a variance of normal prior bias, and x_i and x_j are members of linear regression coefficients, which all are assumed to be also normal priors. We also use a White kernel that represents a white noise given by

$$k\left(x_i, x_j\right) = 1.0 \quad x_i == x_j \text{ else } 0. \tag{2}$$

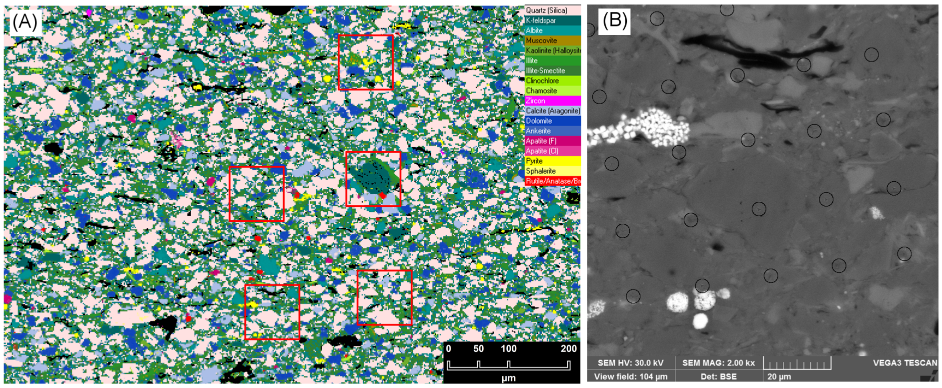

Fig. 1: (A) an example of MAPS image of mineralogical distribution in this study and (B) an SEM image of an indentation area of 25 indentations corresponding to a lower-left red box in (A)

The last regressor is XGBoost (Chen and Guestrin, 2016), which is an ensemble tree-based algorithm. It attempts to predict a dependent variable by combining the estimates of a set of weaker models. We use 100 weaker estimator with maximum tree depth of five, regularization parameter $\lambda = 1.0$, and regularization parameter $\Gamma = 0.0$.

3 RESULTS

3.1. Using mineral composition to predict hardness or reduced Young's modulus

We here use the mineral composition listed in Table 1 as input and either hardness (H) or reduced Young's modulus (E_r) as output to a regressor. In Table 1, there are 17 types of mineral. We use the *unclassified* and *Porosity* labels to represent a part of SEM images that we cannot define and a void inside a porous media, respectively. So, in total, our regressor has an input of 20 features.

As described earlier, we test four types of regressors; 1. linear regression (LR), 2. support vector regression (SVR), 3. Gaussian process regression (GP), and 4. extreme gradient boosting regression (XGBoost). The detailed settings of each model, as well as how we train them, are provided in Section 2. The results of the coefficient of determination R^2 calculated by

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}, \tag{3}$$

where RSS is a sum of squares of residuals, and TSS is a total sum of squares of all regressors, are presented in Table 2. These R^2 results are calculated from the test set, which is randomly selected from the training set (5 % of the training set). From this table, we observe that all models perform poorly (i.e., $R^2 < 0.5$). We will discuss how to improve these results in the following sections.

Table 2: R^2 results of using mineral composition as an input and either hardness (H) or reduced Young's modulus (E_r) as output.

output	R^2			
	LR	SVR	GP	XGBoost
H	0.241	0.347	0.249	0.345
E_r	0.333	0.427	0.355	0.161

3.2. Using mineral composition to cluster the data and hardness to predict reduced Young's modulus

To improve the predictability of E_r, we now first cluster our dataset by mineral composition using BIRCH algorithm (Zhang et al., 1996). We fix *threshold* as 0.001 and use *number of clusters* as hyperparameters. Our results are presented in Table 3. These R^2 results are calculated from the test set, which is randomly selected from the training set (5 % of the training set). Here, we use a number of clusters from 1 to 5 and four types of regressors; 1. linear regression (LR), 2. support vector regression (SVR), 3. Gaussian process regression (GP), and 4. extreme gradient boosting regression (XGBoost). We observe that as the number of clusters increases, the R^2 improves. Besides, these R^2 results are much better than those presented in Table 2. The LR and GP regressors provide the best R^2. However, LR is much computationally cheaper than GP.

Table 3: R^2 results of different models using hardness (H) as input and reduced Young's modulus (E_r) as output. We cluster the data using mineral composition.

number of clusters	R^2			
	LR	SVR	GP	XGBoost
1	0.717	0.722	0.717	0.649
2	0.921	0.906	0.921	0.923
3	0.921	0.906	0.921	0.916
4	0.923	0.915	0.924	0.905
5	0.933	0.925	0.933	0.906

The number training data per cluster is presented in Table 4, and the clustering results presented by box plots for each mineralogy for the number of clusters of 5 in Figure 2. We present box plots only for the number of clusters of 5 because, from Table 3, it shows to deliver the best R^2. We note that Table 3 is constructed from the test set (randomly selected 5 % of the training set) while Table 4 and Figure 2 are constructed from the training set.

From Table 4, the number of members for each cluster represents the number of training data that are classified in each cluster. For example, in the case of the number of clusters is = 1, all the training data, 237 data points, is classified as the first cluster. In the case of the number of clusters = 3, there are 167 data points in the first cluster, 61 data points are in the second cluster, and 9 data points are classified in the third cluster. We can observe that the first cluster of the number of clusters = 4 case is a combined the first and the second clusters of the number of clusters = 5 case. The first cluster of the number of clusters = 3 case is a combined the first and the second clusters of the number of clusters = 4 case. This trend goes on until the number of clusters = 1 case.

Table 4: Clustering results for the training set: we cluster the data using mineral composition.

number of members for each cluster	number of clusters				
	1	2	3	4	5
1	237	x	x	x	x
2	228	9	x	x	x
3	167	61	9	x	x
4	103	64	61	9	x
5	37	66	64	61	9

Figure 2 presents box plots of a composition fraction for each mineralogy presented in Table 1 for each cluster (only for the number of clusters = 5 case). To elaborate Figures 2a-e, represent 1st, 2nd, 3rd, 4th, and 5th cluster, respectively. From these figures, we observe that the 1st, Figure 2a, cluster represents the most heterogeneous material where many minerals co-exists. The 2nd, Figure 2b, cluster is dominated by feldspar while the 3rd, Figure 2c, cluster is predominantly quartz. The 4th, Figure 2d, majority contains smectite. The 5th, Figure 2e, cluster is dominated by dolomite and akerite. We show examples of the SEM images corresponding to each cluster shown in Figure 2.

4 CONCLUSIONS

We want to estimate reduced Young's modulus of highly heterogeneous materials given a set of nanoindentation experiments. We cluster the dataset from mineralogy classification through scanning electron microscope (SEM) images and observe that the number of clusters of five delivers the most accurate results. We also illustrate the mineralogy representing each cluster. Subsequently, we test four types of regression models, namely linear regression, support vector regression, Gaussian process regression, and extreme gradient boosting regression. The R^2 results are not much different, and linear regression and Gaussian process regression provide the most accurate prediction ($R^2 = 0.93$ for the test set). Although further analysis needs to be performed with more data, our work suggests that the balanced iterative reducing and clustering using hierarchies (BIRCH) in conjunction with linear regression or Gaussian process regression can be very accurate to predict the reduced Young's modulus of highly heterogeneous materials from nanoindentation experiments.

5 ACKNOWLEDGMENTS

This work was supported by the Laboratory Directed Research and Development program (project 218328) at Sandia National Laboratories. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

REFERENCES

1. Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

2. Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

3. Lucas, B., Oliver, W., Pharr, G., and Loubet, J. (1996). Time dependent deformation during indentation testing. *MRS Online Proceedings Library (OPL)*, 436:233–238.

4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

5. Yoon, H., Ingraham, M. D., Grigg, J., Rosandick, B., Mozley, P., Rinehart, A., Mook, W. M., and Dewers, T. (2020).

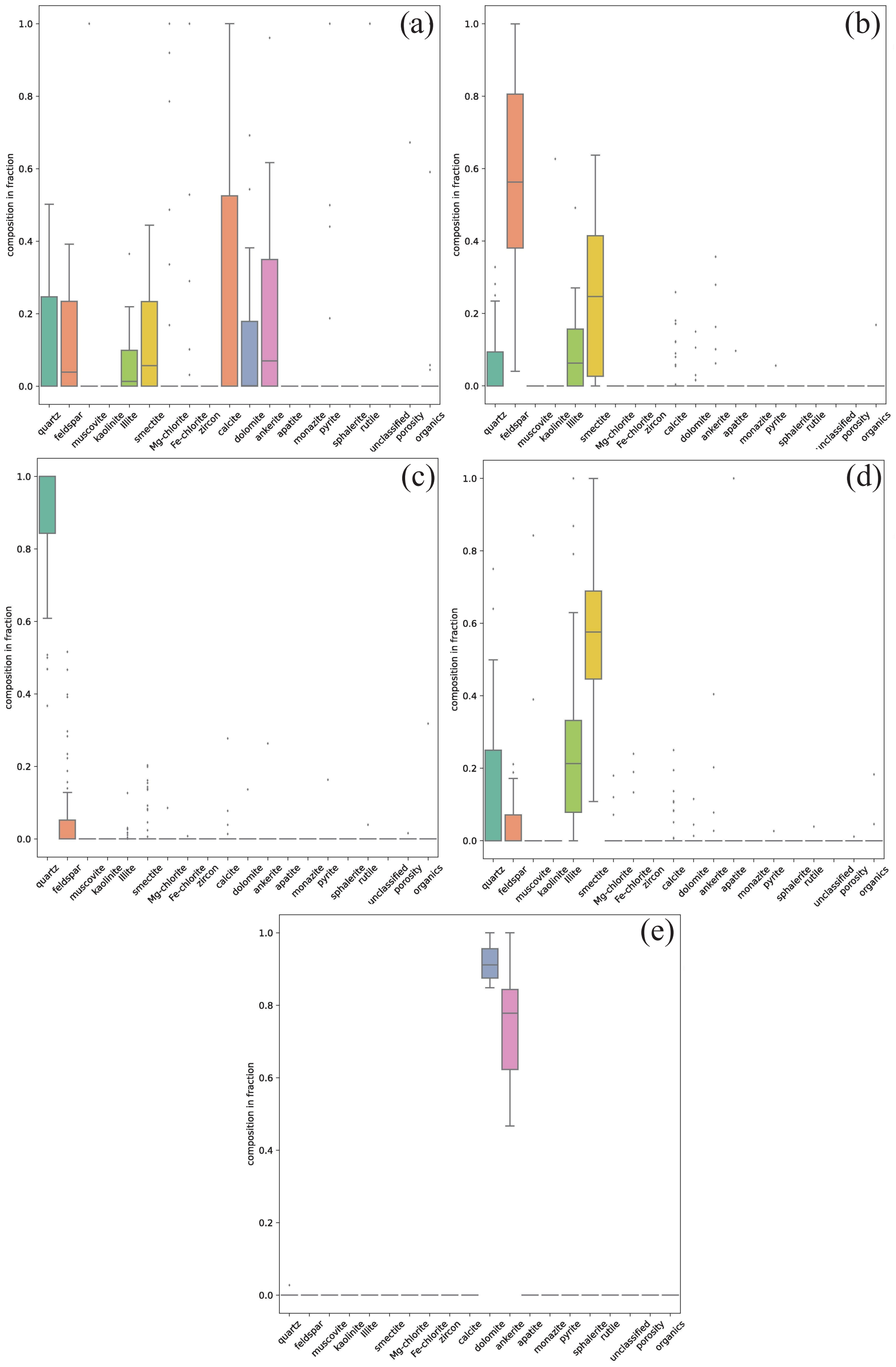

Fig. 2: Clustering results presented by box plots for each mineralogy for the number of clusters of 5. (a) 1st cluster, (b) 2nd cluster, (c) 3rd cluster, (d) 4th cluster, (e) 5th cluster. We cluster the data using mineral composition.

Impact of Depositional and Diagenetic Heterogeneity on Multiscale Mechanical Behavior of Mancos Shale, New Mexico and Utah, USA. In *Mudstone Diagenesis: Research Perspectives for Shale Hydrocarbon Reservoirs, Seals, and Source Rocks*, pages 121–148. The American Association of Petroleum Geologists.

6. Zhang, T., Ramakrishnan, R., and Livny, M. (1996). Birch: an efficient data clustering method for very large databases. *ACM sigmod record*, 25(2):103–114.

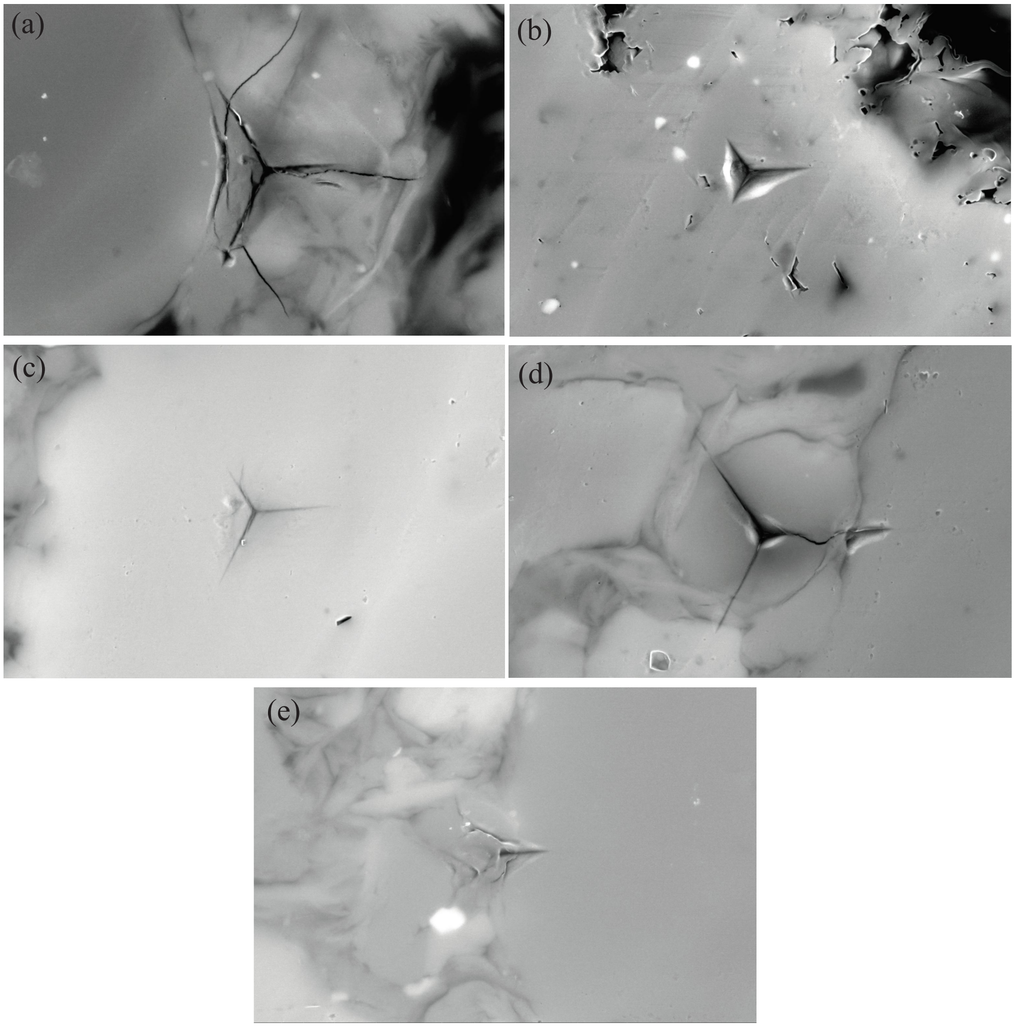

Fig. 3: Examples of scanning electron microscope (SEM) images of (a) 1st cluster, (b) 2nd cluster, (c) 3rd cluster, (d) 4th cluster, (e) 5th cluster corresponding to each cluster shown in Figure 2