1. COVER PAGE

Grant Number:
AFC518-40

Title:
A Smart Device for Mine Dust Characterization and Coal Workers' Health Improvement: Combining Non-Destructive, Element-Specific X-ray CT with Big Data Analytics & Machine Learning

Organization:
Virginia Polytechnic Institute and State University (Virginia Tech)

Principal Investigator:
Cheng Chen, Ph.D., Virginia Tech

Co-Principal Investigators:
Emily Sarver, Ph.D., Virginia Tech
Kray Luxbacher, Ph.D., Virginia Tech
Guohua Cao, Ph.D., Virginia Tech

Contact Information:
Phone: 540-231-2554; Email: chen08@vt.edu

Period of Performance:
August 1, 2017 - May 31, 2019

Table of Content

2. Executive Summary

The primary mission of this project is to develop hardware and software solutions to determine element distribution in the 3D structural space within a microparticle, which has important implications to improving coal worker health and safety with respect to respirable mine dust problems. The specific project goals are:

1) Develop 3D non-destructive, element-specific CT capabilities to identify element heterogeneity in a single mine dust particle at the spatial resolution of 50 nm/pixel.
2) Develop autonomous image pattern recognition capability to extract information for dust particles, such as size, shape, density, and element.
3) Identify potentially predictive correlations between particle size, shape, and chemical composition using machine learning and big data analytics.
4) Develop advanced numerical modeling capabilities to improve the understanding of microparticle transport and deposition (aerodynamic properties) in the human lung.

In this project, "element heterogeneity" refers to variable element distribution throughout the 3D space within a microparticle. For the *first goal*, we developed an artificial intelligence (AI) based imaging and segmentation technology to evaluate element distribution within the 3D structural space of a microparticle. In this technology, Nano-computed tomography (CT) and scanning electron microscope (SEM) are used to scan the same area of microparticle surface, in order to collect training data that contain the correlations between the greyscale CT values and the SEM element information. In the training process, not only the contrast in greyscale CT values is used as a training feature, but also the geometrical characteristics of the interfaces between elements are extracted for training. Next, a random decision forest training and classification process is performed to segment the greyscale CT pixels throughout the entire 3D structural space within the microparticle. In this study, we use 200 decision trees in the random decision forest model, and the final decision is made by voting. The AI model was used to segment element distribution in two custom-made microparticles, and the evaluated surface coating thicknesses were in good agreement with the values provided by the manufacturer. For the *second goal*, the AI capability was incorporated into an in-house image processing and analysis software, DNA-Viz. For the *third goal*, various machine learning (ML) models have been tested for classification of dust particles. First, the k-means clustering, an unsupervised ML method, was used to conduct preliminary data classification. Next, the k-nearest neighbors (k-NN) and support vector machine (SVM) methods, both of which are supervised ML algorithms, were used to classify dust particles based on 100,000 particle data points which have been labeled. We found that the SVM method provides an overall training and testing accuracy about 10% higher than the k-NN, because the SVM mitigates the overfitting issue better. In addition, the SVM model accounts for the geometric property of particles, which implies that there may be underlying correlations between particle geometry and chemical composition. For the *fourth goal*, we conducted fundamental fluid dynamics and particle transport simulations at the pore scale in a synthesized human lung model. The air flow was simulated using the lattice Boltzmann (LB) method, which is a numerical model for solving air flow at the pore scale. Dust particle migration in the human lung model was then simulated using the particle tracking method based on the LB-simulated air flow field. Three particle transport and filtration mechanisms were accounted for in particle tracking, including Brownian motion, streamline advection, and gravitational settling. The simulated dust particle deposition amount was plotted as a function of dust particle size, and the plot showed a "U" shape, which is consistent with the classic theoretical prediction. In these LB simulations, we generated synthesized human lung models

having varying pore size distributions to study their influence on the dust particle deposition amount.

Specifically, previous studies suggested that surface coating of quartz particles may modify the biologically available surface area of quartz particles, which suggests that the silicon element may have less toxicity to the human lung when it is present inside a microparticle than being on the particle surface. Therefore, identification of the 3D spatial distribution of a specific element in a microparticle has important implications to the advanced understanding of the relationship between silica-rich microparticles and occupational respiratory illnesses such as coal workers' pneumoconiosis (CWP) and silicosis. Element distribution information in the 3D structural space of a microparticle will provide a means to assess whether silica particles are pure or coated, which can provide helpful information in identifying the most harmful dust constituents. However, detection and identification of 3D element distribution in a microscopic particle is challenging due to the small spatial scale. This project aimed to integrate Nano-CT and SEM scanning with AI processing to solve the problem. Two custom-made microparticles, in which the 3D element distributions were known, were used to validate the analysis and data processing technology that has been developed during the research period. Specifically, the "single-blind" experimental paradigm was used, which means that the manufacturer of the microparticles gave us accurate element information for only one microparticle, whereas for the other microparticle only rough estimation was provided. The purpose of this "single-blind" experiment was to ensure independent measurements using the developed CT-SEM-AI technology. The results showed that the CT-SEM-AI technology was able to accurately identify the 3D element distributions in both microparticles (please see support letter from Cospheric LLC in Appendices).

3. Concept Formulation and Mission Statement

The primary mission of this project is to develop hardware and software solutions to determine element distribution in the 3D structural space within a microparticle, which has important implications to improving coal worker health and safety with respect to respirable mine dust problems (IARC 1997; ISO 1995; OSHA 2010; WHO 1999; CDC 2006; Castranova et al., 2000; Laney et al., 2012; Suarthana et al., 2011; Laney and Attfield, 2014; Pollock et al., 2010; Sellaro et al., 2015; Johann-Essex et al., 2017). Previous studies suggested that surface coating of quartz particles may modify the biologically available surface area of quartz particles (Harrison et al., 1997). This implies that the silicon element may demonstrate different levels of toxicity to the human lung when it is inside a microparticle versus when it is coated on the microparticle surface. Thus, element distribution information in the 3D structural space of a microparticle will provide a way to assess whether silica particles are pure or coated, which can provide helpful information in identifying the most harmful dust constituents.

The specific research goals in this project are:

1) Develop 3D non-destructive, element-specific CT capabilities to identify element heterogeneity in a single mine dust particle at the spatial resolution of 50 nm/pixel.
2) Develop autonomous image pattern recognition capability to extract information for dust particles, such as size, shape, density, and element.
3) Identify potentially predictive correlations between particle size, shape, and chemical composition using machine learning and big data analytics.
4) Develop advanced numerical modeling capabilities to improve the understanding of mine dust transport and deposition (aerodynamic properties) in the human lung.

Rationale: Specifically, to tackle the first goal, we initially planned to develop the 3D X-ray element imaging capabilities using synchrotron-based X-ray beams. In summer 2018, the two custom-made particles were initially scanned using Argonne National Laboratory's high-resolution Transmission X-ray Microscopy located in the 2-BM beamline (https://www.aps.anl.gov/Imaging). The major challenge was the highly limited access to synchrotron-based X-ray facilities. Therefore, after preliminary analysis and technology development based on synchrotron X-ray, we decided to combine available commercial imaging hardware instruments (Nano-CT and SEM) with AI data processing methods, which we believed is a more cost-effective approach. Specifically, we used the Zeiss UltraXRM-L200 Nano-CT as a non-invasive imaging method to obtain the 3D greyscale value distribution throughout the inside of the microparticle. The UltraXRM-L200 Nano-CT is a non-invasive, 3D imaging method with the highest spatial resolution of 16 nm per pixel length. The acquired greyscale CT values are proportional to the atomic numbers of the elements within the microparticle (Chen, 2016), which suggests that a denser material will have a higher greyscale value in the CT picture (i.e., a higher brightness in the CT image). However, the brightness contrast information provided by 3D Nano-CT does not give direct element "labels". Therefore, in this study, the TESCAN MIRA3 SEM instrument was used to provide element feature (i.e., labels) for the reconstructed 3D Nano-CT images. An AI-based data analytics method, which is based on the random decision tree method, was developed to correlate the Nano-CT information to the SEM information. In this way, it becomes possible to identify the 3D spatial distribution of a specific element of interest inside a microparticle, which has important implications to the advanced understanding of the relationship between silica-rich microparticles and occupational respiratory illnesses such as coal CWP and silicosis because existing studies imply that the silicon element may demonstrate

different levels of toxicity to the human lung when it is inside a microparticle versus when it is coated on the microparticle surface. **Figure 1** illustrates the hardware and software components of the developed technology.

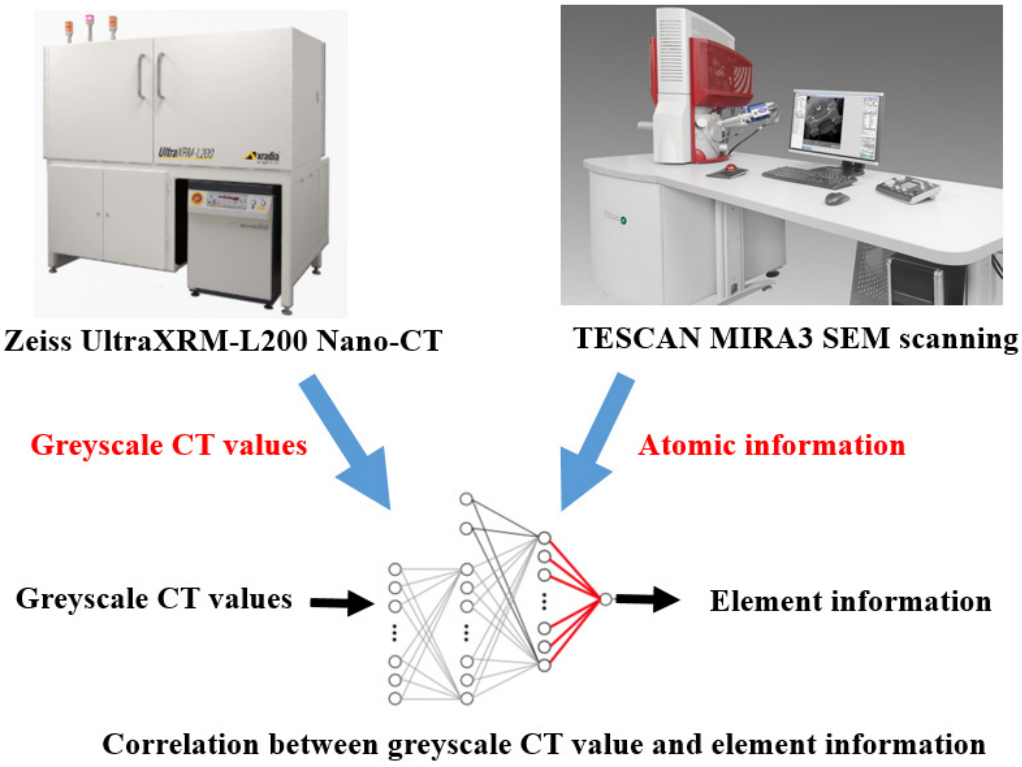


Figure 1. Schematic workflow demonstrating the hardware/software-integrated technology. Specifically, the Zeiss UltraXRM-L200 3D Nano-CT is used to obtain the 3D spatial distribution of greyscale values within a single microparticle, whereas the TESCAN MIRA3 SEM scanning is used to provide element "labels". The AI data processing method, which is based on the random decision tree method in this project, is used to correlate the 3D greyscale values to the element information.

4. Proof-of-Concept Technology Components

In this proof-of-concept project, the first research component aims to develop 3D non-destructive, element-specific CT capabilities to identify element heterogeneity in a single mine dust particle at the spatial resolution of 50 nm/pixel. Based on the first component, the second research component aims to develop autonomous image pattern recognition capabilities to extract information for dust particles, such as size, shape, density, and element. The third and fourth research components are relatively separate from the first two. Specifically, the third research component aims to identify potentially predictive correlations between particle size, shape, and chemical composition using machine learning and big data analytics methods, based on 100,000 microparticle data points which have been labeled. The fourth research component aims to develop advanced numerical modeling capabilities to improve the fundamental understanding of mine dust transport and deposition (aerodynamic properties) in the human lung.

4.1. *Goal 1*: Develop imaging and AI methods to identify element heterogeneity in a single mine dust particle at the spatial resolution of 50 nm/pixel

For Goal 1, we developed an artificial intelligence (AI) based imaging and segmentation technology to evaluate element distribution within the 3D structural space of a microparticle. In this technology, Nano-computed tomography (CT) and scanning electron microscope (SEM) are used to scan the same area of microparticle surface, in order to collect training data that contain the correlations between the greyscale CT values and the SEM element information. In the training process, not only the contrast in greyscale CT values is used as a training feature, but also the geometrical characteristics of the interfaces between elements are extracted for training. Next, a random decision forest training and classification process is performed to segment the greyscale CT pixels throughout the entire 3D structural space within the microparticle. In this study, we use 200 decision trees in the random decision forest model, and the final decision is made by voting. The AI model was used to segment element distribution in two custom-made microparticles, and the evaluated surface coating thicknesses were in good agreement with the values provided by the manufacturer.

We ordered two types of custom-made microparticles that had heterogeneous mineral distributions in the 3D structural space. The *purpose* was to use microparticle samples that had well-controlled 3D element and mineral distributions so that they can be used to calibrate and validate the developed imaging and AI tools. Here, “well-controlled” means that the manufacturer of these microparticles knows the detailed elemental and geometric information of the particles, which can be used as the “ground truth” to validate the developed SEM-CT-AI integrated technology. Specifically, the developed imaging and AI tools will be used to identify element distribution in the microparticles and to determine the thickness of the surface coating layers.

Figure 2 illustrated the 2D cross sections for the structures of the two 3D microparticles that were custom-made by Cosheric LLC (please see Cospheric in References). The first microparticle has a barium titanate glass core and is coated by aluminum. The second microparticle has a soda lime glass core and is coated by silver. These two well-controlled microparticles were designed to account for the following two scenarios: in the first scenario a lighter mineral is coated on the surface of a heavier microparticle, whereas in the second scenario a heavier mineral is coated on the surface of a lighter microparticle. We designed these two custom-made, well controlled microparticles to account for the mining health scenarios where the elements of interest are lighter and denser than the micropaticle core materials separately.

The surface coating thicknesses, 375 nm for the aluminum-coated microparticle and 100 nm for the silver-coated microparticle, were reported by the manufacturer when we placed the ordered (email records available upon request). However, after imaging analysis and AI segmentation, we found that the surface coating thickness for the aluminum coating was 288 nm and for the silver coating it was 416 nm. It is obvious that there was a noticeable difference between manufacturer-reported silver coating thickness (100 nm) and our AI-measured silver coating thickness (416 nm). Therefore, we contacted the manufacturer to discuss this. It turned out that the silver coating thickness of 100 nm was based on their guess without rigorous measurements. The manufacturer then used a rigorous laboratory method to calculate the silver coating thickness. Specifically, the manufacturer calculated the surface coating thickness by analyzing the difference in true particle density before and after surface coating. They used a helium gas pycnometer which measures all of the microparticle volume that is impenetrable by helium, and then measured the total microparticle mass on an ultra-precision balance; the mass and volume information was then used to calculate the true particle density. Using this true-particle-density method, the manufacturer found that the silver coating thickness was 435 nm, which was very close to our AI-based measurement (416 nm). Please see the support letter from the microparticle manufacturer, Cospheric LLC, attached in Section 7 – Appendices.

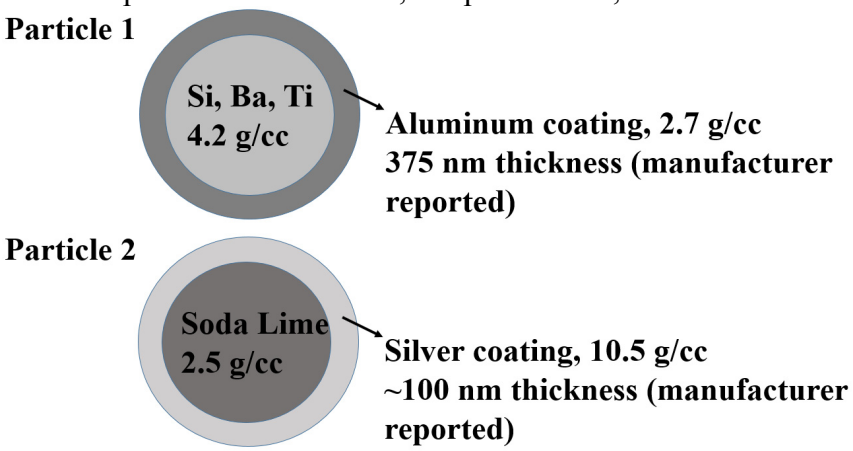


Figure 2. 2D cross sections of two custom-made, heterogeneous 3D microparticles with surface coatings. The first microparticle has a barium titanate glass core and is coated by aluminum. The second microparticle has a soda lime glass core and is coated by silver. The microparticles were custom-made by Cospheric LLC (see Cospheric in References). These two well-controlled microparticles were designed to account for the following two scenarios: in the first scenario a lighter mineral is coated on the surface of a heavier microparticle, whereas in the second scenario a heavier mineral is coated on the surface of a lighter microparticle.

Figure 3 displays the aluminum-coated microparticles and the silver-coated microparticles in the laboratory. Both particles are fine-sized. The silver-coated microparticles have diameter between 45 and 53 μm with a mean diameter of 49 μm, whereas the aluminum-coated microparticles have diameter between 30 and 100 μm with a mean diameter of 60 μm. The manufacturer, Cospheric LLC, reported that the surface coating thickness was roughly 100 nm (without rigorous measurements in the lab) for the silver-coated microparticles and 375 nm (based on true particle density analysis in the lab) for the aluminum-coated

microparticles. **Table 1** illustrates the detailed microsphere information provided by the manufacturer.

Figure 3. Aluminum-coated microparticles (left) and silver-coated microparticles (right). Both particles are fine-sized. The silver-coated microparticles have diameter between 45 and 53 μm with a mean diameter of 49 μm, whereas the aluminum-coated microparticles have diameter between 30 and 100 μm with a mean diameter of 60 μm.

Table 1. Information of the two types of custom-made microparticles.

Sample	Name	Characterization
A	Silver-coated soda lime glass microparticles	45-53 μm particle diameter About 100-nm silver coating thickness reported by the manufacturer (without rigorous measurements in the lab)
B	Aluminum-coated barium titanate glass microparticles	30-100 μm particle diameter 375-nm aluminum coating thickness reported by the manufacturer (based on true particle density analysis in the lab)

4.1.1. Nano-CT analysis

The instrument used in the Nano-CT scanning test was UltraXRM-L200, which is designed with the ability to visualize the samples in 3D space with internal structures and features. This CT scanner provides non-destructive 3D resolution up to 16 nm per pixel length with fully automated data acquisition.

High-resolution images can be obtained from both low- and high-atomic-number materials, composites, polymers, and biological samples without the addition of contrasting

agents. The associated software allows the calculation of pore size, density, and standard surface metrology. There are two types of instrument scanning modes in this Nano-CT. The first mode is the large field of view (LFOV) mode, where the field of view is 64 μm and the resolution is 64 nm per pixel length. The second mode is high resolution (HR) mode, where the field of view is 16 μm and the resolution is 16 nm per pixel length. In this project, we use the LFOV mode to scan the two microparticles in order to enlarge the field of view to visualize the entire microparticles.

4.1.2. SEM Analysis

SEM images were obtained using the TESCAN MIR3 XMH device to characterize the surface coating of the microparticles. MIRA3 is a high-performance SEM system which features a high brightness Schottky emitter to achieve high-resolution and low-noise images. MIRA3 offers all the advantages that come with the newest technologies and developments in SEM, which include an ultra-fast scanning system, faster image acquisition, dynamic and static compensation, and built-in scripting for user-defined applications. The highest resolution of this device can reach 1 nm per pixel length, with the highest potential of 20 kV.

4.1.3. XRD Analysis

X-ray diffraction (XRD) analyses were conducted using the Rigaku Ultima IV device to obtain the diffraction angles of various materials in the coated microparticles. The Ultima IV incorporates Rigaku's patented cross beam optics (CBO) technology for permanently mounted, permanently aligned, and user-selectable parallel and focusing geometries. The Ultima IV X-ray diffractometer can perform many different measurements, and also incorporates fully automatic alignments. When coupled with CBO, the automatic alignment capability makes the Ultima IV X-ray diffractometer the most flexible system available for multipurpose applications.

4.1.4. XRF Analysis

The X-ray fluorescence (XRF) analysis was conducted using Rigaku Supermini 200 to determine the chemical composition of the coated microparticles. Because the XRD analysis is a semi-quantitative analysis, the accurate amount of each component obtained via XRD should be quantified using the XRF result. The Supermini200 is a compact, benchtop wavelength dispersive X-ray fluorescence (WDXRF) for elemental analysis. It has several advantages, including light element sensitivity, exceptional elemental resolving power, and low limits of detection. The XRF experiment begins with exposing a sample to high-energy photons from an X-ray tube, which induces transitions of electrons between atomic orbitals and results in the emission of fluorescent photons. By measuring the energy and intensity (count rate) of these photons, qualitative and quantitative information for the elemental composition can be obtained.

4.1.5. Artificial Intelligence Segmentation

After the high-resolution Nano-CT scanning of these two microparticles, the 3D CT images will be analyzed and segmented for pore-space and solid-phase to build the 3D structural models. The mineral of each pixel can be identified according to the attenuation coefficient value in the CT images. The effective values of the attenuation coefficients for these minerals were calculated via the intensity of X-ray transmission through the grain at the

projection direction versus the length of transmission. The intensity was taken from the projection images, and the transmission length was measured in reconstructed slices. For each pixel unit along the x axis (the direction of X-ray), it is possible to acquire intensity I(x) and the transmission length h(x). On the other side of the pixel we have the information of I(h). According to the equation of attenuation determination (Equqation1), we obtain the attenuation coefficient, μ, which is calculated through the linear regression analysis of ln[I(h)/I(0)] versus h.

$$I_h = I_0 e^{\mu h} \quad (1)$$

With all the μ value calculated, the 3D greyscale CT values can be divided into several groups according to the SEM and XRD results. The percentage of each group was then further verified by the XRF result. Based on this workflow, the 3D mapping of mineral groups can be achieved using all the information acquired above.

An in-house trainable segmentation software was developed to improve the segmentation efficiency and accuracy. A machine learning algorithm was developed for image processing, where two or more classes of the images will be defined manually for training. The feature of selected input image of the class will be extracted and converted to a set of vectors of float values. In the training process, not only the contrast in greyscale CT values is used as a training feature, but also the geometrical characteristics of the interfaces between elements are extracted for training. If the training ends correctly, the 3D images will be completely segmented for simulations.

To segment the grayscale 3D images from Nano-CT scanning, a random forest training and classification process is performed. To address this trainable segmentation process, we need to prepare a grayscale 2D image with N pixels as the input training data and a few image filters as the feature set. Because the silver-coated microsphere in the 3D X-ray images has a sharp interface with the void space, we readily located the 2D cross section of the 3D image as the input training data which matches the 2D image from SEM analysis by comparing radii of their enclosed disks. The latter forms a mapping m from any training data subset D to the triple-element set of segmentation, i.e. m: D -> {carbon, aluminum, void space}, serving as the supervisor of the training. Among a wide range of image filters, we selected five commonly used filters as the feature set, which are the Gaussian blur, Sobel, Hessian, difference of Gaussians, and membrane projections. The filters applied to the training data will help capture the features of the data and build correlations, which we call decision trees, between the grayscale 3D CT image and its corresponding segments.

In the training stage, we stochastically generated 200 data subsets (i.e., decision trees) with the capacity of M (M < N, i.e. M = N / 20) pixels out of the underlying input training data by using the bootstrap sampling method. A decision tree is produced using two randomly chosen filters from the feature set, namely F = {f1, f2}, and one of the data subsets, namely P = {$pixel_i$ | i = 1, ..., M}. A node of a tree owns two properties: the property "set" is the subset of P to be trained for the classification; the property "class" is the classified segment if the node is a leaf and is NULL otherwise.

We adopted the Classification and Regression Tree (CART) algorithm to branch the tree nodes, where the concepts of Gini value and Gini index are used. Let f be any filter, D be any data subset and |f(D)| be the cardinal number of f(D), then D can be separated into |f(D)| ordered subsets with ascending order in the values of f(D), such that the subset D_k correspond to the k^{th} value of f(D). Likewise, m(D_k) can be separated into three subsets: D_{k1},

D_{k2} and D_{k3}. The Gini value of the k^{th} subset D_k of D filtered by *f*, reflecting its purity with the triple-element segmentation, is defined as:

$$\mathrm{Gini}(D_k) = 1 - \frac{D_{k1}^2 + D_{k2}^2 + D_{k3}^2}{D_k^2}. \quad (2)$$

The corresponding Gini index is defined as:

$$\mathrm{Gini}_{\mathrm{ind}}(f, D) = \sum_{k=1}^{|f(D)|} \mathrm{Gini}(D_k) \frac{|D_k|}{|D|}. \quad (3)$$

In the classification stage, a generated decision tree was able to classify any grayscale CT data set, in 2D or 3D space, into three segments according to its own strategy. A single strategy was likely biased, which means it performs well for one case but not for another scenario. Therefore, diversity was desired to satisfy the ergodicity requirement and to enhance segmentation robustness. The 200 data subset along with the feature set of filters produced 200 decision trees. These trees with all individual strategies comprised a forest for the 3D image segmentation. By applying the trained forest with diversity to each pixel on each layer of the grayscale 3D CT images, we obtained 200 decisions regarding classification and the final decision is made by voting.

The following C-style pseudo code demonstrates the algorithm of a decision tree's recursive producing process in the AI-based segmentation process:

```
void produceTree(P,F)
{
    newNode = NULL;
    if (all pixels in P are of segment s)
    {
        newNode.set = P;
        newNode.class = s;
        exit;
    }
    if (F = NULL || F(P) is a single value set)
    {
        newNode.set = P;
        newNode.class = the dominant segment of P;
        exit;
    }
    f = argmin( Gini_ind(f1,P), Gini_ind(f2,P) );
    separate P into |f(P)| subsets P1, P2, ..., P|f(P)|;
    Pk = argmin( Gini(P1), Gini(P2), ..., Gini(P|f(P)|) );
    // generate two nodes
    node[0].set = P1 ∪ P2 ∪ ... ∪ Pk;
    produceTree(node[0].set,F-{f});
    node[1].set = Pk+1 ∪ Pk+2 ∪ ... ∪ P|f(P)|;
    if node[1].set == NULL
        node[1].class = the dominant segment of P;
    else
        produceTree(node[1].set,F-{f});
}
```

Figure 4. The C-style pseudo code that demonstrates the algorithm of a decision tree's recursive producing process in the AI-based segmentation process.

4.2. *Goal 2*: Develop autonomous image pattern recognition capability to extract information for dust particles, such as size, shape, density, and element.

For Goal 2, the autonomous image pattern recognition capability has been developed and incorporated in the software package, DNA-Viz, as illustrated in Figure 9. Correlations between particle geometry and chemical composition will be discussed in the machine learning task in Section 4.3. In the future, we aim to incorporate the AI capabilities, which relate microparticle geometry to chemical information, into the image processing software.

4.3. *Goal 3*: Identify potentially predictive correlations between particle size, shape, and chemical composition using machine learning and big data analytics.

In this task, we developed machine learning (ML) tools to classify respirable microparticles, based on 100,000 particle data points that have been labeled by co-PI Sarver's team (Sellaro et al., 2015; Johann-Essex et al., 2017) using advanced SEM analyses in a separate project. These labeled microparticle data will be used as the training and testing data sets in the process of ML model development. It is widely known that purely data-driven ML technologies require a large amount of labeled data, and a big challenge is that the labeling and classification of samples are expensive. Therefore, these 100,000 labeled particle data points provide valuable information for the development of the ML models in this project.

Specifically, various ML models have been tested for the classification of microparticles. First, the *k*-means clustering, an unsupervised ML method, was used to conduct preliminary data classification. Next, the k-nearest neighbors (k-NN) and support vector machine (SVM) methods, both of which are supervised ML algorithms, were used to classify dust particles based on the 100,000 particle data points which have been labeled. We found that the SVM method provides an overall training and testing accuracy about 10% higher than the k-NN, because the SVM mitigates the overfitting issue better. In addition, the SVM model accounts for the geometric property of particles, which implies that there may be underlying correlations between particle geometry and chemical composition.

Figure 5 demonstrates the schematic workflow for the ML procedure used in this project. The section of data and methods includes all the data backgrounds and data processing methods. The data sources, background, and the four data preprocessing methods will be described. After preprocessing, we selected the *k*-means clustering method to conduct preliminary data examination. We then chose two different ML models to test the accuracy and precision of this model. These two models are the KNN method and the SVM method. Using the results from the SVM model, we made the mismatch discussion to improve the model. After model improvement, we implemented data visualization for the geometry-chemical correlation. Based on the visualization result, we expressed the correlation in a mathematical way.

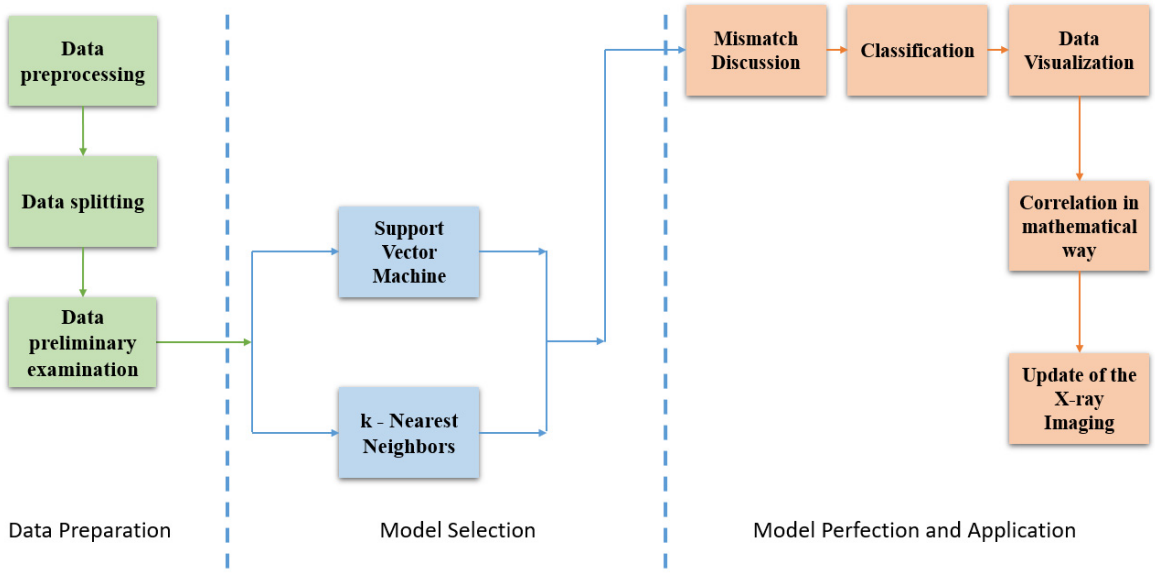


Figure 5. Overall workflow for microparticle classification and for investigating the correlations between particle size, shape, and chemical composition using ML tools.

4.3.1 Data Preprocessing

The data preprocessing step includes data splitting, feature subset selection, oversampling, and normalization. Data splitting is the process of partitioning available data into two portions, usually for cross-validation purpose (Picard and Kenneth, 1990). One portion of the data is used to develop a predictive model, and the other to evaluate the model performance. As an essential part of data preprocessing, we split the data according to their natural properties. The first splitting strategy is to split data by their regions and locations. We can also split the data based on their mineral compositions. This splitting strategy is more straightforward and convenient than the first strategy. The splitting categories include alumino-silicate, carbonaceous, carbonate, mixed carbonaceous, quartz, aluminum, iron, titanium and some unclassified data.

Feature subset selection is one of the approaches to reduce data dimensionality. This selection aims to avoid the redundant and irrelevant features in the data set. The redundant feature is mainly duplicate information which is contained in one or more other attributes. The irrelevant features contain no information that is useful for the data mining task. To avoid the redundant and irrelevant features described above, we selected eight chemical elements from the total 48 features in the original data set, which are directly related to the particle mineral composition and uncorrelated with each other, as the features to fit the model in this project.

Oversampling is a technique in data analytics and ML to adjust the class distribution of a data set. One important purpose of using the oversampling technique is to correct a bias in the original data set. There are many simple and complex oversampling techniques, including the synthetic minority over-sampling technique (Chawla et al. 2002) and the ADASYN algorithm (He et al., 2008). In our project, we enlarged the data sets of some specific classes that had relatively low data volumes to obtain higher classification precision.

Data normalization is essential for model fitting. There are various types of data normalization methods, such as the min-max normalization, decimal scaling normalization, and

Z-score normalization. In this project, we selected the Z-score normalization method for our data set. The formula is written as:

$$v' = \frac{v - \mu_A}{\sigma_A} \quad (4)$$

where v is the original data point, μ_A is the mean data value, σ_A is the standard deviation, and v' is the data point after normalization.

4.3.2. Preliminary Data Examination

First, we used unsupervised learning to perform the preliminary data examination. The goal of this examination is to label the unclassified data. The amount of unclassified data is 2552. The *k*-means clustering algorithm was used as the unsupervised learning algorithm, which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster having the nearest mean (Hartigan and Manchek, 1979). We then developed a model with the *k*-means algorithm to check the silhouette value (SV) of these data. The SV is a measure of how similar an object is to its own cluster compared to other clusters (Rousseeuw, 1987). The SV ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to the neighboring clusters. If most objects have a high value of SV, then the clustering configuration is appropriate. On the other hand, if many points have a low or negative SV, then the clustering configuration may have too many or too few clusters. We evaluated the SV with different class numbers to determine the optimal class number.

4.3.3. ML Model Selection

The target of the data-driven approach is to classify the particles from SEM-EDX images automatically. The first model we selected to classify the data is based on the *k*-nearest neighbors (*k*-NN) algorithm. The *k*-NN algorithm is a type of instance-based learning, where the function is only approximated locally and all computation is deferred until classification (Altman, 1992). The *k*-NN algorithm is a simple yet surprisingly efficient algorithm, which may be among the simplest of all ML algorithms. Moreover, it is an instance-based classifier, which means we can use the observations directly. In the *k*-NN ML model, we chose eight chemical elements as the features (inputs) and the particle classes as the labels (outputs). The eight chemical elements are O, Al, Si, C, Mg, Ca, Ti, and Fe. After feature selection, the next step is cross-validation. The cross-validation is to assess how the results of a statistical analysis will generalize to an independent data set (Kohavi, 1995). The *k*-fold cross-validation, especially the 10-fold cross-validation method, is commonly used. In the *k*-fold cross-validation process, the original sample is randomly partitioned into *k* equal-sized subsamples. Only one subsample (the test data set) will be retained to test the other *k* - 1 subsamples, and the other *k*-1 subsamples are used as the training data set. Therefore, we split all the data into 10 folders for cross-validation to check the error and to conduct fine tuning of the parameters. After finishing the cross-validation, we split data by mineral compositions into nine categories (including the unclassified data) and then developed a model using the *k*-NN algorithm. We then calculated the accuracy from the confusion matrix based on the *k*-NN algorithm to assess the model performance.

Although the *k*-NN algorithm is simple and convenient to use, it cannot solve the issue of overfitting in many cases. Considering oversampling and feature dimension reduction, the better choice for avoiding overfitting is to change the model. The second model we tested is based on the support-vector machine (SVM) algorithm. The SVM method is one of the supervised learning models with associated learning algorithms that analyze data for classification and

regression analyses (Cortes and Vladimir, 1995). A model based on the SVM algorithm is a depiction of the data in a manner of scattering plot. Therefore, the data points within separate groups can be divided by a clear boundary line. The SVM algorithm has its unique advantage to avoid overfitting. The SVM algorithm only needs to consider the number of support vectors. However, for other ML algorithms, we need to consider the data dimension, which is why the possibility of overfitting in the SVM algorithm is relatively low.

4.3.4. Classification and the Geometry-Chemical Correlation

We fitted the SVM model and then expressed the particle geometry-chemical correlation in a mathematical way. Consequently, when a new, unknown data point comes in, we can predict which class it belongs to by using the geometry-chemical correlation.

The multiclass classification method that we used is one-vs-one (OvO) in SVM. In OvO reduction, one trains K(K−1)/2 binary classifiers (rules) for a K-way multiclass problem. To finish the prediction, all classifiers (rules) are applied to a new, unknown data sample. If one class receives the highest number of "+1" predictions, the new sample will fall into this class. The expression for the classification is shown in Equation 5:

$$Y = sign\big(f(X)\big) = \begin{cases} +1 & \text{Class 1} \\ -1 & \text{Class 2} \end{cases} \quad (5)$$

where X is the normalized data point with n features. The function $f(X)$ is calculated as:

$$f(X) = \sum_{i=1}^{n} w_i x_i + \mathrm{b} \quad (6)$$

where n is the total feature number, w_i is the weighting coefficient, and b is the bias coefficient. w_i and b can both be calculated from the SVM model. In this project, we combined several classes and then had five classes in total. These five classes are quartz, alumino-silicate, heavy minerals, carbonate, and carbonaceous. Because the class number is 5, the total classifier (rule) number is 10. **Table 2** demonstrates the comparison classifier (rule) matrix.

Table 2. Comparison rule matrix for the geometry-chemical correlation prediction

	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Rule 6	Rule 7	Rule 8	Rule 9	Rule 10
Quartz	1	1	1	1	0	0	0	0	0	0
Alumino-silicate	-1	0	0	0	1	1	1	0	0	0
Heavy Mineral	0	-1	0	0	-1	0	0	1	1	0
Carbonate	0	0	-1	0	0	-1	0	-1	0	1
Carbonaceous	0	0	0	-1	0	0	-1	0	-1	-1

Figure 6 illustrates the schematic workflow of class predication for new particle data. Based on Table 2, Equation 5, and Equation 6, we can predict the class of a new data point using the workflow demonstrated in Figure 6.

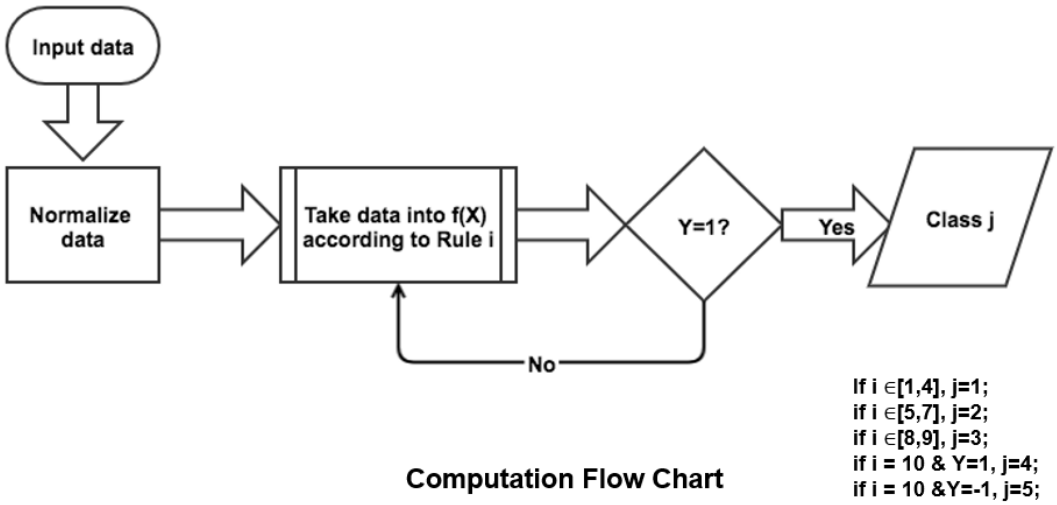


Figure 6. Workflow of particle class predication based on the SVM ML model. The ML model input, $f(x)$, depends on the eight chemical features and one geometric feature, as shown in Equations 5 and 6.

4.4. *Goal 4*: Develop advanced numerical modeling capabilities to improve the fundamental understanding of microparticle transport and deposition (aerodynamic properties) in the human lung.

In this task, we conducted fundamental fluid dynamics and particle transport simulations at the pore scale in a synthesized human lung model. The air flow was simulated using the lattice Boltzmann (LB) method (Chen and Doolen, 2998; Succi et al., 1991; Succi, 2001; Chen et al., 2008, 2009, 2010, 2013, 2016; Chen and Zhang, 2009), which is a numerical model for solving fluid flow at the pore scale. Dust particle migration in the human lung model was then simulated using the particle tracking method based on the LB-simulated air flow field. Three particle transport and filtration mechanisms were accounted for in particle tracking, including Brownian motion, streamline advection, and gravitational settling. The simulated dust particle deposition amount was plotted as a function of dust particle size, and the plot showed a "U" shape, which is consistent with other lung model predictions (ICRP, 1994; Choi and Kim, 2007; Hofmann, 1982; Asgharian et al., 2001; Koblinger and Hofmann, 1990; and Hofmann, 2011). In these LB simulations, we generated synthesized human lung models having varying pore size distributions to study their influence on the dust particle deposition amount.

4.4.1. Overview of the Mathematical and Numerical Approaches

In this task, the Particle Flow Code 3D (PFC3D) (Itasca, 2008) was used to generate 3D pores having various pore size heterogeneity (i.e., varying pore size distributions) in a simplified human lung model. An in-house numerical code (Fan et al., 2018) was developed to discretize the pore structure of the human lung model and to import it into the lattice Boltzmann (LB) simulator as internal boundary conditions of air flow modeling to simulate air flow in the pore spaces. Microparticle transport and deposition in the human lung pore space were then numerically simulated at the pore scale based on the LB-simulated air flow field, and three transport mechanisms that regulate fine particle migration, including interception collection, Brownian

motion, and gravitational settling, were accounted for. Using this numerical workflow, one can obtain the pore structure geometry and pore-scale flow characteristics to study fine particle migration and deposition throughout the 3D pore space. The numerical results of pore-scale particle tracking and deposition will be fitted using a continuum-scale fine particle deposition model to determine the macroscopic deposition coefficient, as well as how fine particle size and pore size heterogeneity influence the macroscopic deposition coefficient. The following sections provide details of the mathematic and numerical methods.

4.4.2. Continuum-Scale Mathematical Model for Microparticle Deposition and Empirical Correlation for the Deposition Coefficient

When microparticles migrate through the pore space, they can attach on solid surfaces and clog the pore spaces. The mass of microparticle removed from the air flow is equal to the mass of deposited microparticles on solid surfaces, as illustrated in Equation 7:

$$U \cdot dC \cdot A_{cs} \cdot dt = -\alpha \cdot (A_{cs} \cdot dx \cdot \phi \cdot C) \cdot dt \tag{7}$$

where C is microparticle concentration in the air (kg/m^3); U is flow velocity (m/s); A_{cs} is the cross section area perpendicular to the main air flow direction (m^2); x is the distance along the air flow direction (m); t is time (s); ϕ is porosity of the human lung model (i.e., ratio of total void space volume to the total lung volume); α is the rate of microparticle attachment on the lung inner wall surfaces, which indicates the fraction of suspended microparticles that attach onto inner wall surfaces per unit time (s^{-1}). After rearrangement of equation 7, one obtains:

$$\frac{dC}{C} = -\alpha\phi\frac{dx}{U} \tag{8}$$

By integrating Equation 8, microparticle concentration can be written as:

$$C = be^{-\frac{\alpha\phi}{U}x} \tag{9}$$

where b is a constant and equal to the influent microparticle concentration, C_0 (kg/m^3). Equation 9 is thus re-written as:

$$C = C_0 e^{-\frac{\alpha\phi}{U}x} \tag{10}$$

By defining the deposition coefficient (m^{-1}), $\lambda = \alpha\phi / U$, one can calculate microparticle concentration in the air along the main flow (x) direction, which is consistent with the classical colloid filtration model (Yao et al., 1971):

$$C = C_0 e^{-\lambda x} \tag{11}$$

As deposited micrpparticles accumulate within the lung model, the lung inner wall surfaces are coated with attached microparticles. The analysis of microparticle deposition mass considers a mass balance between microparticle concentration in the air flow (kg/m^3), C, and deposited microparticle concentration on lung inner wall surfaces (kg/m^3), σ, which indicates the mass of deposited microparticles per unit volume of the lung model. This mass balance expression is written as:

$$U \cdot dC \cdot A_{cs} \cdot dt = -A_{cs} \cdot dx \cdot d\sigma \tag{12}$$

After rearrangement, one obtains

$$\frac{\partial \sigma}{\partial t} = -U \frac{\partial C}{\partial x} \tag{13}$$

By substituting Equation 11 into Equation 13, one obtains

$$\sigma = tUC_0 \lambda e^{-\lambda x} + b_2 \tag{14}$$

where b_2 is a constant and t is time. When x=0 and t=0, particle deposition mass is equal to 0, which suggests that $\sigma(x=0,t=0)=b_2=0$. Therefore, Equation 14 is re-written as:

$$\sigma(x,t) = tUC_0 \lambda e^{-\lambda x} \tag{15}$$

The clean bed deposition coefficient, λ, can also be determined using an empirical correlation given by McDowell (1986):

$$\lambda = \frac{3}{2}\frac{(1-\phi)}{d_m}\left[4A_s^{1/3}\left(\frac{Ud_m}{D}\right)^{-2/3} + 0.56A_s\left(\frac{A}{\mu d_p^2 U}\right)^{1/8}\left(\frac{d_p}{d_m}\right)^{15/8} + 2.4\times10^{-3}A_s\left(\frac{v_s}{U}\right)^{1.2}\left(\frac{d_p}{d_m}\right)^{-0.4}\right] \tag{16}$$

where d_m is the grain diameter of the porous medium (m), d_p is the microparticle diameter (m), A is the Hamaker's constant with a typical value in the range of 10^{-13} to 10^{-12} erg. A_s is the dimensionless Happel correction factor that accounts for the influence of the neighboring spheric particles and the pore geometry effect, written as:

$$A_s = \frac{1-\varphi^5}{1-1.5\varphi+1.5\varphi^5-\varphi^6} \tag{17}$$

where $\varphi = (1-\phi)^{1/3}$.

In this empirical correlation for the macroscopic deposition coefficient (Equation 16), microparticle transport is influenced by the mechanisms of Brownian motion, interception collection, and gravitational settling, and these effects are additive. The first term within the brackets of Equation 16 accounts for the Brownian motion mechanism, the second term represents the interception collection mechanism, and the third term represents the gravitational settling mechanism.

4.4.3. Pore-Scale Particle Tracking and Deposition Based on Pore Air Flow Field

This section will describe the numerical method for tracking pore-scale microparticle movement in the air flow and deposition in the lung pore space based on the LB-simulated air flow field. Details about the LB method will be given in a later section. The numerical results of pore-scale particle tracking and deposition will be fitted using the continuum-scale mathematical model of microparticle deposition described previously to determine the macroscopic deposition coefficient, as well as how microparticle size and lung pore size heterogeneity influence the macroscopic deposition coefficient.

Equation 18 illustrates the microparticle tracking algorithm based on the LB-simulated air flow field:

$$\boldsymbol{x}(t+\Delta t) = \boldsymbol{x}(t) + \boldsymbol{v}(\boldsymbol{x}(t))\bullet\Delta t + \boldsymbol{v}_s(\boldsymbol{x}(t))\Delta t + \sqrt{6D\Delta t}\boldsymbol{\xi} \quad (18)$$

where $\boldsymbol{x}$ is the vector indicating the position of the microparticle in the 3D lung pore space (m); t is time (s); Δt is the time step used in particle tracking (s); $\boldsymbol{v}$ is the pore air flow velocity vector which is determined by the pore-scale LB simulation (m/s); $\boldsymbol{v}_s$ is the Stokes settling velocity in air (m/s); D is the diffusivity of the microparticles in air (m^2/s); $\boldsymbol{\xi}$ is a random vector, of which the direction is uniformly distributed in the 3D space and the magnitude is a random variable having zero mean and unit variance.

The Stokes settling velocity in air is calculated using Equation 19:

$$v_s = \frac{2}{9}\frac{(\rho_p - \rho_f)}{\mu} g r^2 \quad (19)$$

where μ is the dynamic viscosity of air (kg/m/s); g is the gravitational acceleration (m/s^2); ρ_p is the mass density of the microparticle (kg/m^3); ρ_f is the mass density of air (kg/m^3). The diffusivity of the microparticle in air is calculated using the Stokes-Einstein equation:

$$D = \frac{k_b T}{6\pi r \mu} \quad (20)$$

where k_b is the Boltzmann constant and equal to 1.38×10^{-23} m^2kg/s^2k^{-1}; T is the absolute temperature (k); r is the microparticle radius (m).

Using this particle tracking method, the displacement of an individual microparticle in the pore space of the human lung model is determined by convection, gravitational settling, and Brownian motion, which correspond to the second, third, and fourth terms on the right hand side of Equation 16, respectively. When an individual microparticle follows the convective streamline and collide with the lung inner wall surface, this deposition mechanism is referred to as interception collection as described by the second term on the right hand side of Equation 16. Gravity can cause a microparticle to deviate from the streamline and move downward, which may lead to microparticle collision and deposition on the lung inner wall surface; this deposition mechanism is referred to as gravitational settling as described by the third term on the right hand side of Equation 16. In addition, random Brownian motion causes a microparticle to deviate from the convective streamline as well, which can also lead to collision and deposition on the lung inner wall surfaces, as described by the fourth term on the right hand side of Equation 16. It should be noted that all these three mechanisms contribute to the transport and deposition of a microparticle in the pore space of the lung model. For relatively small particles, the Brownian motion mechanism dominates, whereas for relatively large particles the gravitational settling mechanism dominates.

In this project, the Monte Carlo (MC) method was used to simulate the transport and deposition of 10,000 microparticles in the pore space of the human lung model based on LB-simulated pore air flow field using Equation 18. When a microparticle collides with the inner wall surface in the human lung model, the longitudinal (x) position where the collision occurs is recorded. At the end of the Monte Carlo simulation, the distribution of all the collision positions in the x direction is fitted using the macroscopic fine particle deposition model (Equation 15), which determines the continuum-scale deposition coefficient, λ. In this project, a microparticle that attaches on the inner wall surface will not detach, which suggests that the sticking coefficient is equal to one. In addition,

in this project the fine particle size is much smaller than the aveage pore size in the human lung model and it is assumed that the deposit mass is relatively low, which suggests that the fitted deposition coefficient is close to the clean bed deposition coefficient.

4.4.4. Discretization of the Pore Structure in the 3D Human Lung Models

3D pore structures of the human lung model were discretized, extracted, and then imported into the LB model as internal boundary conditions of air flow modeling to simulate pore-scale, single phase air flows in the pore spaces. Specifically, the 3D lung pore geometry was discretized using a 3D mesh grid having a resolution of 0.05 mm/pixel in the x-, y-, and z-directions. Becasue the average pore diameter in the lung model is close to 1 mm, which is 20 times larger than the single pixel size, the human lung geometry is well resolved with this resolution.

4.4.5. Lattice Boltzmann Method for Air Flow Simulation in the 3D Human Lung Model

In this project, the LB method is used to simulate pore-scale air flow fields in the pore spaces of a human lung model, which are then used to track microparticle transport and deposition using the algorithm illustrated in Equation 18. The LB method is a numerical method for solving the Navier-Stokes equations and based on microscopic physical models and mesoscale kinetic equations. In comparison with conventional fluid dynamic models, the LB method has many advantages. For example, it is explicit in evolution equation, simple to implement, natural to parallelize, and easy to incorporate new physics such as interactions at fluid-solid interface.

The LB simulator used in this study has been validated by direct comparisons with analytical solutions and laboratory measurements in the PI Chen's previous works. It was then optimized with high-performance graphics processing unit (GPU) parallel computing, which enhances the computational speed by a factor of 1,000 and led to an in-house LB code, GPU-enhanced lattice Boltzmann simulator (GELBS). In this work, the D3Q19 lattice structure (19 velocity vectors in 3D space) was used because of its advantage in keeping a good balance between computational stability and efficiency.

Particle distribution in the Bhatnagar-Gross-Krook (BGK)-based, single-relaxation-time LB equation is given by

$$f_i(\mathbf{x}+\mathbf{e}_i\,\Delta t, t+\Delta t) = f_i(\mathbf{x},t) - \frac{f_i(\mathbf{x},t) - f_i^{eq}(\rho,\mathbf{u})}{\tau}, \qquad (i = 0,1,2\ldots18) \tag{21}$$

where $f_i(\mathbf{x},t)$ is the particle-distribution function specifying the probability that fluid particles at lattice location $\mathbf{x}$ and time t travel along the ith direction; $\mathbf{e}_i$ is the lattice velocity vector corresponding to direction i, defined as:

$$\boldsymbol{e}_i = c\begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 & 0 & 1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 & 1 & -1 & 1 & -1 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \end{bmatrix}$$

where $c = \Delta x / \Delta t$, in which Δx is the lattice spacing and Δt is the time step; τ is the dimensionless relaxation time related to kinematic viscosity by $\upsilon = (2\tau - 1)\Delta x^2 / 6\Delta t$; $f_i^{eq}(\rho, \mathbf{u})$ is the equilibrium distribution function selected to recover the macroscopic Navier-Stokes equations and given by

$$f_i^{eq}(\rho,\mathbf{u}) = \omega_i \rho\left[1 + \frac{3\mathbf{e}_i\cdot\mathbf{u}}{c^2} + \frac{9(\mathbf{e}_i\cdot\mathbf{u})^2}{2c^4} - \frac{3\mathbf{u}^2}{2c^2}\right] \tag{22}$$

where w_i is the weight coefficient calculated as:

$$\omega_i = \begin{cases} 1/3 & i = 0 \\ 1/18 & i = 1...6 \\ 1/36 & i = 7...18 \end{cases}$$

The macroscopic fluid density and velocity are calculated with the following two equations:

$$\rho = \sum_{i=0}^{18} f_i \tag{23}$$

and

$$\mathbf{u} = \frac{\sum_{i=0}^{18} f_i e_i}{\rho} \tag{24}$$

Air pressure is calculated using $p = c_s^2 \rho$, where c_s is the speed of sound. In the LB *D3Q19* model, $c_s^2 = c^2/3$.

In practice, two-relaxation-time and multi-relaxation-time LB schemes have been developed to mitigate numerical instability in simulating high-Reynolds-number flows and avoid nonlinear dependency of numerical error on fluid viscosity (Li and Huang, 2008; Ginzburg, 2008; Ginzburg et al., 2010). In this study, we replaced the BGK-based collision operator with a two-relaxation-time collision operator and selected the optimal combination of the symmetric and asymmetric eigenfunctions in order to reduce numerical errors resulting from the bounce-back boundary condition.

For air flow numerical simulation, we imposed a periodic boundary condition with a constant pressure difference, ΔP, in the longitudinal direction and no-slip boundary conditions on the four lateral sides and interior solid surfaces. More details about the LB simulator and associated GPU optimization can be found in our previous papers (Chen et al., 2016).

5. Proof of Concept Evaluation

5.1. *Goal 1*: Develop imaging and AI methods to identify element heterogeneity in a single mine dust particle at the spatial resolution of 50 nm/pixel

5.1.1. Nano-CT and SEM Results

The silver-coated microparticles were analyzed using SEM scanning. **Figure 7** illustrates The SEM image at 500 times of magnification. The microparticle diameters vary from 45 to 55 µm. It is observed that the silver was coated smoothly and uniformly on the surface of the microspheres. However, some defects were observed on the surface of the microparticles when the magnification increased to 3,000 times, as demonstrated in **Figure 8**.

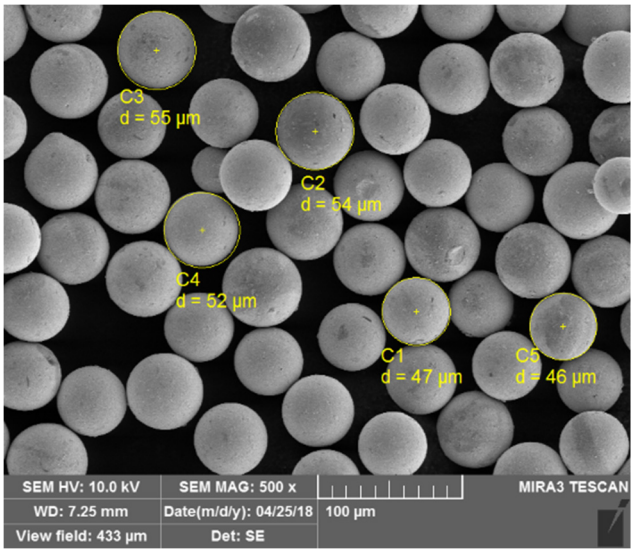


Figure 7. SEM image of the silver-coated microparticles at 500 times of magnification. The microparticle diameters vary from 45 to 55 µm. At this resolution it appears that the silver was coated smoothly and uniformly on the surface of these microspheres.

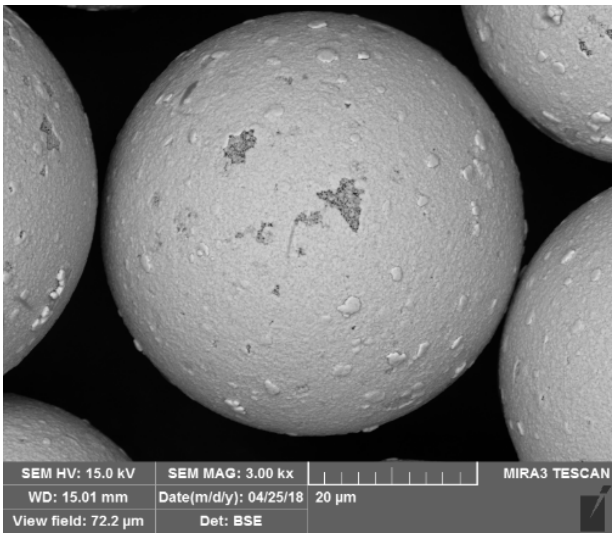


Figure 8. SEM image of the silver-coated microparticles at 3,000 times of magnification. Uneven coating of silver can be observed at this spatial resolution.

On the other hand, the aluminum-coated microparticles were examined using XRF and XRD analyses. The microspheres were compressed into a pellet with a diameter around 1.5 inch and the pellet was then placed inside the device to conduct XRF analysis. **Table 3** shows the XRF analysis results. Because the microparticle core was made of silica and barium titantate, the corresponding formula was used for further processing. In summary, we found approximately 16.52 mol% of aluminum, 5.68 mol% of silica, 68.89 mol% of barium titantate, and 8.91 mol% of titanium oxide. The result of XRD analysis is also critical to the subsequent AI analysis and will be used as supporting information in the AI segmentation.

Table 3. XRF Analysis of the aluminum-coated microspheres.

No.	Component	Percentage	Series	Intensity
1	Na	0.36	Na-KA	0.0318
2	Mg	0.0165	Mg-KA	0.004
3	Al	19.7	Al-KA	24.4993
4	Si	3.16	Si-KA	3.0167
5	S	0.0068	S-KA	0.0142
6	Cl	0.064	Cl-KA	0.1836
7	K	0.13	K-KA	0.172
8	Ca	3.68	Ca-KA	7.3687
9	Ti	23.3	Ti-KA	10.1737
10	Fe	0.0808	Fe-KA	0.076
11	Ni	0.0988	Ni-KA	0.1699
12	Sr	1.07	Sr-KA	8.7193
13	Ba	48.3	Ba-LA	6.1658

5.1.2. AI Segmentation Result

After the Nano-CT scanning, the 3D CT image datasets were processed with various adjusted parameters to obtain nearly 1000 images in the three (x, y, z) principal directions. An in-house software, DNA-Viz, was developed to process the grayscale CT images and to conduct other calculations, as demonstrated in **Figure 9**. The image was processed with the AI function to distinguish various minerals in the microparticle. The 3D structure of the microparticle was then reconstructed.

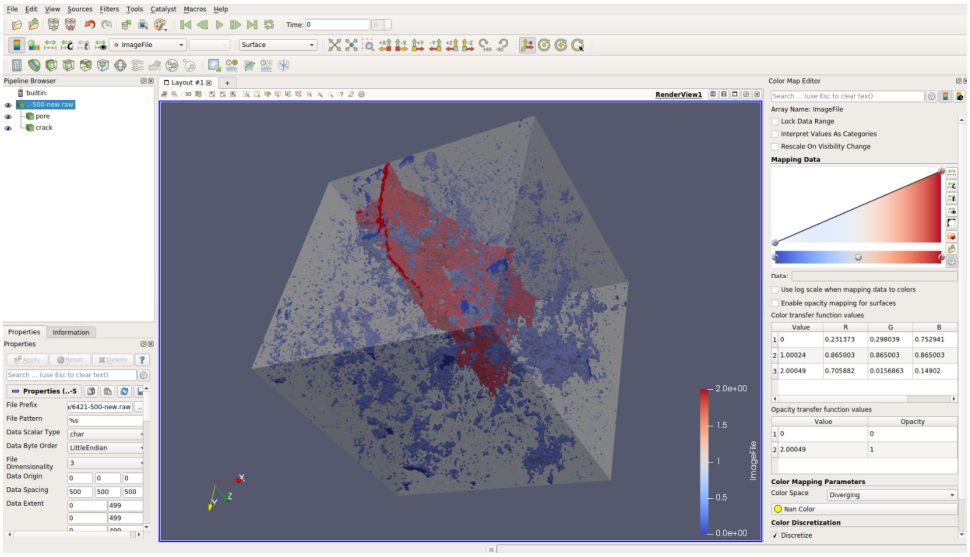


Figure 9. User interface of the in-house image processing and calculation software, DNA-Viz.

Figure 10 illustrates the 3D images of a silver-coated microparticle before and after the AI recognition. The 2D images in the XY, XZ, and YZ planes are then presented in **Figure 11**. A bright circle can be clearly observed, and it was classified by the AI function as silver because it has the highest greyscale value in the Nano-CT scanning. The portion inside the silver coating was classified as the core glass material of the microparticle. The area outside the microsphere was classified as the void space. Most of the silver layer has uniform thickness. However, there were surfaces that did not have any silver coating, as illustrated by the red circle in Figure 11, and surfaces that had extra silver deposition thickness, as shown by the yellow circle in Figure 11. The average thickness of the silver coating was 6-7 pixels, suggesting that the silver coating had thickness of 384 to 448 nm. It is clear that there was a noticeable difference between manufacturer-reported silver coating thickness (~100 nm) and our AI-measured silver coating thickness. Therefore, we contacted the manufacturer, and it turned out that the silver coating thickness of 100 nm was based on their guess without rigorous measurements. The manufacturer then used a rigorous laboratory method to calculate the silver coating thickness. Specifically, the manufacturer calculated the surface coating thickness by analyzing the difference in true particle density before and after surface coating. They used a helium gas pycnometer which measures all of the microparticle volume that is impenetrable by helium, and then measured the total microparticle mass on an ultra-precision balance; the mass and volume information was then used to calculate the true particle density. Using this true-particle-density method, the manufacturer found that the silver coating thickness was 435 nm, which was very close to the average coating thickness (416 nm) from our AI-based measurement. Please see the supporting letter from the microparticle manufacturer, Cospheric LLC, attached in Section 7 – Appendices.

In addition, based on direct voxel counting, there were 12,323,776 silver voxels and 234,748,256 glass voxels in the Nano-CT images of the microparticle. Assuming that the glass core and the silver-coated-microparticle are both perfect spheres, the glass core has a radius of 24.49 μm and the silver-coated microparticle has a radius of 24.91 μm, leading to a

silver coating thickness of 420 nm, which agrees well with the measurements from the AI classification.

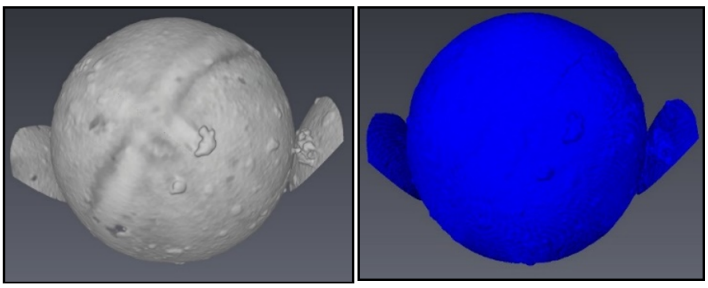


Figure 10. 3D structure of a silver-coated microparticle before (left panel) and after (right panel) AI recognition. The digital image resolution is 64 nm per pixel length.

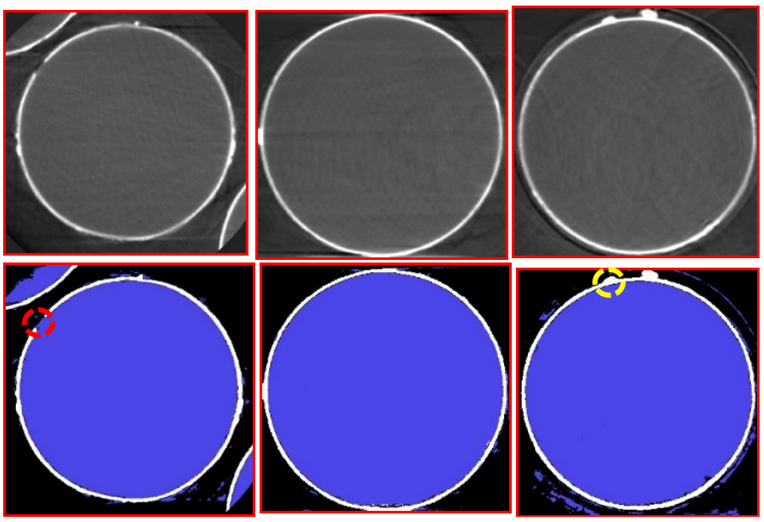


Figure 11. 2D images of a silver-coated microparticle in the XY, XZ, and YZ planes before (top row) and after (bottom row) AI recognition. The digital image resolution is 64 nm per pixel length.

An aluminum-coated microparticle was analyzed using the Nano-CT at the same resolution (64 nm per pixel length). Because the size of this microparticle was larger than the silver-coated microparticle, the acquisition of the full microparticle was impossible with the current field of view. Therefore, only a portion of the aluminum-coated microparticle was scanned to evaluate the coating material and thickness. 3D grayscale CT images of the aluminum-coated particle before and after AI recognition are presented in **Figure 12**. Because of the poor X-ray absorption property of aluminum, the surface coating was difficult to see and distinguish. However, we still observed that the aluminum coating was highly porous and the coating surface had high roughness with "spikes" geometry, as demonstrated in **Figure 13**. From the 2D images, the AI function estimated that the average

thickness of the aluminum coating was 4-5 pixels, which was 256 to 320 nm and close to the value (375 nm) provided by the manufacturer.

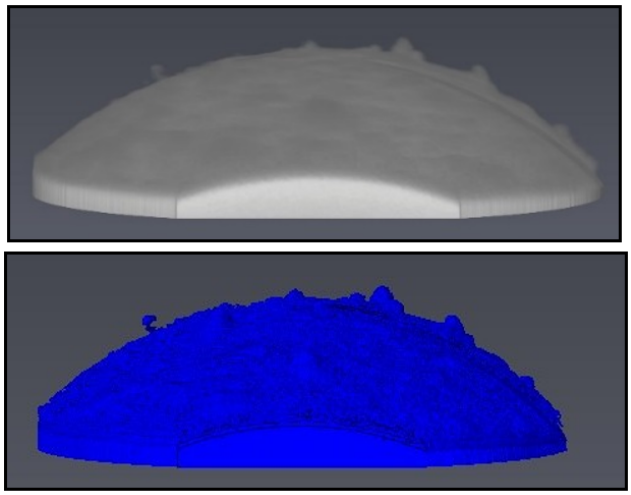


Figure 12. 3D structure of a portion of an aluminum-coated microparticle before (top) and after (bottom) AI recognition. It is clear that the aluminum coating layer has high surface roughness, which is consistent with the Nano-CT raw images. The digital image resolution is 64 nm per pixel length.

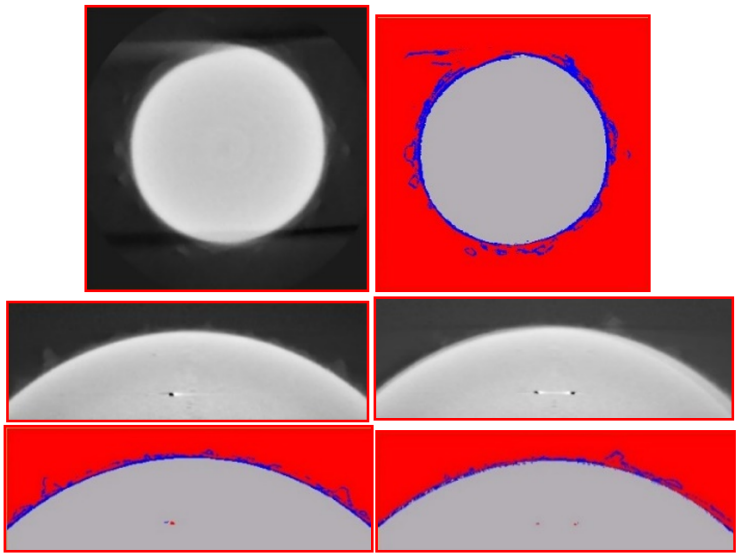


Figure 13. 2D images of an aluminum-coated particle in the XY, XZ, and YZ planes before (grayscale images) and after (color images) AI recognition. It can be observed that the aluminum coating was highly porous and the coating surface had high roughness. The digital image resolution is 64 nm per pixel length.

5.1.3. Combination of Nano-CT and Confocal Micro-X-ray Fluorescence

The team also collaborated with the Los Alamos National Laboratory (LANL) to test a direct way of visualizing 3D element distribution in the microparticles, on the basis of the combination of Nano-CT scanning and the confocal micro-X-ray fluorescence (MXRF). During the Nano-CT scan, the HR mode of the Nano-CT was used to scan an aluminum-coated microparticle, leading to a spatial resolution of 16 nm per pixel length, as illustrated in **Figure 14**. The Nano-CT scanning at the 16 nm resolution confirmed the AI-based estimate of the aluminum coating thickness. Also, the 16 nm resolution confirmed that the aluminum coating was highly open and porous, which was consistent with our CT scanning using the LFOV mode as shown in Figure 12. Next, an aluminum-coated microparticle was placed in the confocal MXRF for element distribution analysis as demonstrated in **Figure 15**. The source and optic provide a beam about 50 -100 μm in diameter. The detector and optic only detect the fluorescent X-ray from this volume. It is a 3D spatially confined beam. Elements are identified and mapped. **Figure 16** illustrates the schematic plot of the equipment setup for the confocal MXRF.

Figure 14. 2D cross sections of 3D Nano-CT scanning of an aluminum-coated microparticle at the LANL. These 2D images clearly show the porous and open structure of surface aluminum coating, which is consistent with our finding in Nano-CT scanning The digital image resolution is 16 nm per pixel length using the HR mode of the Nano-CT scanner.

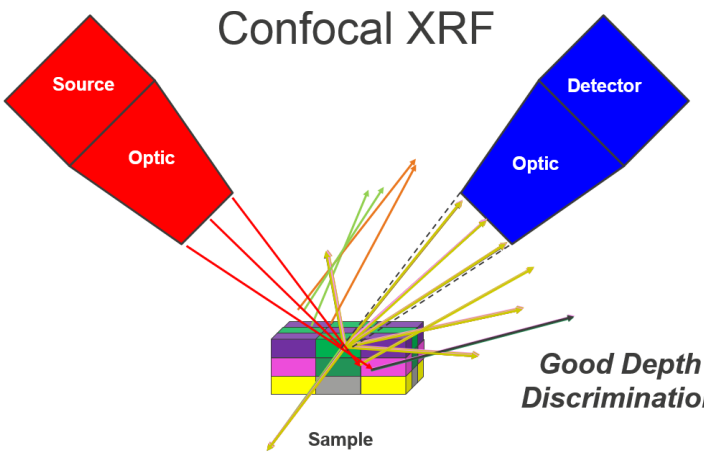


Figure 15. Source and optic provide a beam about 50 -100 microns in diameter. Detector and optic only detect the fluorescent X-ray from this volume. It is a 3D spatially confined beam. Elements are identified and mapped.

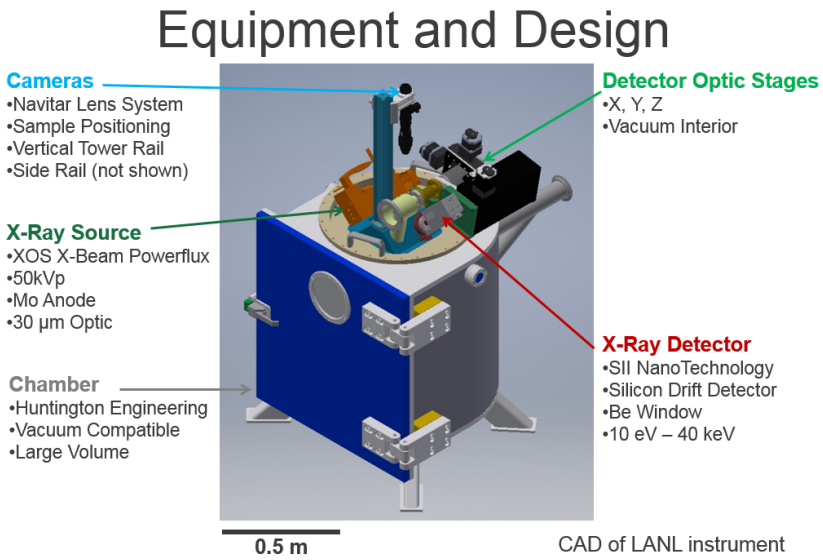


Figure 16. Schematic plot of the equipment setup for the confocal MXRF.

Figure 17 illustrates the signal counts of barium and aluminum from the confocal MXRF scanning. It is clear that the counts of barium were much higher than aluminum, because aluminum coating was thin and highly porous on the microparticle surface.

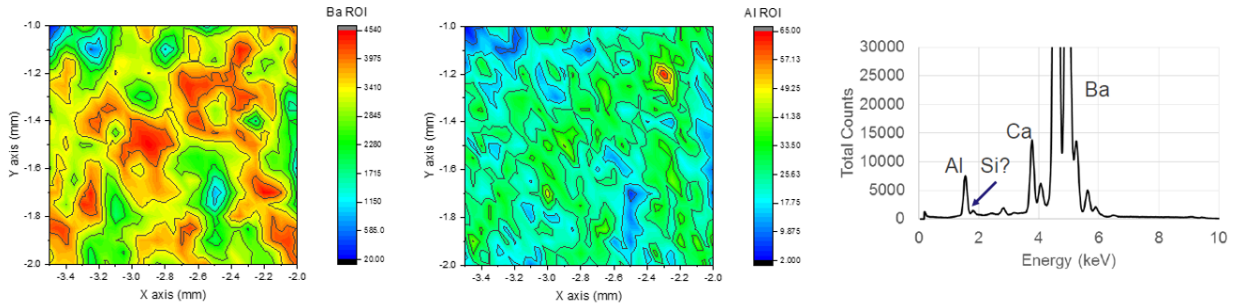


Figure 17. Signal counts for barium and aluminum from the confocal MXRF.

5.1.4. Discussion and Future Research Directions

In this research task, the AI model was used to segment element distribution in two custom-made microparticles, and the element analysis results and evaluated surface coating thicknesses were in good agreement with the values provided by the manufacturer. Therefore, we think that the AI-based element segmentation method is promising and works well in some specific scenarios. However, more fundamental research is needed along this direction. In order to make this AI-based method to work, in the ML training process, not only the contrast in grayscale CT values is used as a training feature but also the geometrical characteristics of the interfaces between minerals are extracted for training. This suggests that we will need at least two or three minerals present at the microparticle surface so that the SEM scanning can see them and label them as the ML features (i.e., the ML model inputs). These features not only include the contrast in grayscale CT values but also account for the geometrical characteristics of the interfaces between minerals.

In this research task, we had been using the LFOV mode of the Nano-CT scanner, which gives a spatial resolution of 64 nm per pixel length. Note that this Nano-CT scanner can achieve the highest spatial resolution of 16 nm per pixel length if the HR mode is used in the Nano-CT scanning; in this case, associated with the SEM analysis and AI-based mineral segmentation, it is possible to resolve a mineral aggregate having a 1D size of 32 nm (if we define that the minimum object size is two times of the CT pixel size). In this project, we chose to use the LFOV model, which gives a resolution of 64 nm per pixel length, because the diameters of the two custom-made microparticles are in the range of 40-50 μm. The resolution of 64 nm per pixel length gives us larger fields of view when we scan the two microparticles. If smaller surface-coated microparticles are available (e.g., with diameters smaller than 10 microns), then the HR mode can be used, which will lead to a spatial resolution of 16 nm per pixel length.

5.2. *Goal 2*: Develop autonomous image pattern recognition capability to extract information for dust particles, such as size, shape, density, and element.

For Goal 2, the autonomous image pattern recognition capability has been developed and incorporated in the software package, DNA-Viz, as illustrated in Figure 9. The evaluation results about the correlations between particle geometry and chemical composition will be discussed in the machine learning task in Section 5.3. In the future, we aim to incorporate the

AI capabilities, which relate microparticle geometry to chemical information, into the image processing software.

5.3. *Goal 3*: Identify potentially predictive correlations between particle size, shape, and chemical composition using machine learning and big data analytics.

5.3.1. Results and Discussion

The *k*-means clustering method was first used to test the optimal class number for the unclassified microparticles in the original data set. **Figure 18** illustrates the mean SV value for the unclassified microparticles as a function of class number in the preliminary data examination using the *k*-means clustering method. The highest mean SV is 0.6962, which corresponds to the class number of 2. The second highest SV value, 0.6923, is for the class number of 3. The mean SVs from other class numbers are all lower than these two values. Therefore, the unclassified data contain two or three classes. The next step is to classify these data points depending on the chemical elements, shape information, and region locations.

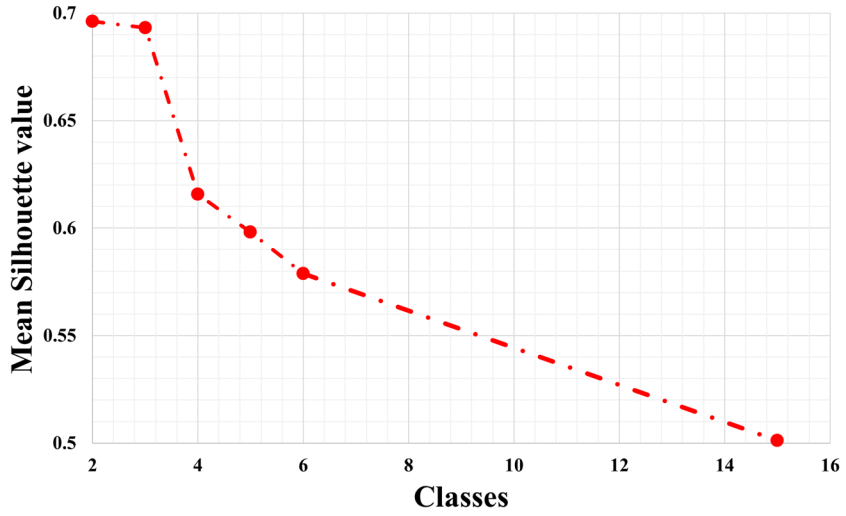


Figure 18. Mean SV as a function of class number in the *k*-means clustering ML model.

The *k*-NN ML model is a voting algorithm. The overall error rate, by comparisons with the original data labels, is 4.2% for the ten-folder cross-validation and 3.08% for the testing data, which suggests a satisfying performance of the *k*-NN model. We then checked the error distribution of the k-NN model. **Figure 19** illustrates the error distribution of the k-NN model over the eight mineral categories. We found that the *k*-NN ML model had the highest prediction error rates in the categories of heavy mineral: aluminum, heavy mineral: titanium, and heavy mineral: iron, which were 51.5%, 22.0%, and 8.5%, respectively. It is worth mentioning that the data amount of these three categories are 27, 46, and 670, which are far less than the other categories. To reduce the model prediction errors in these three categories, we used the data oversampling method.

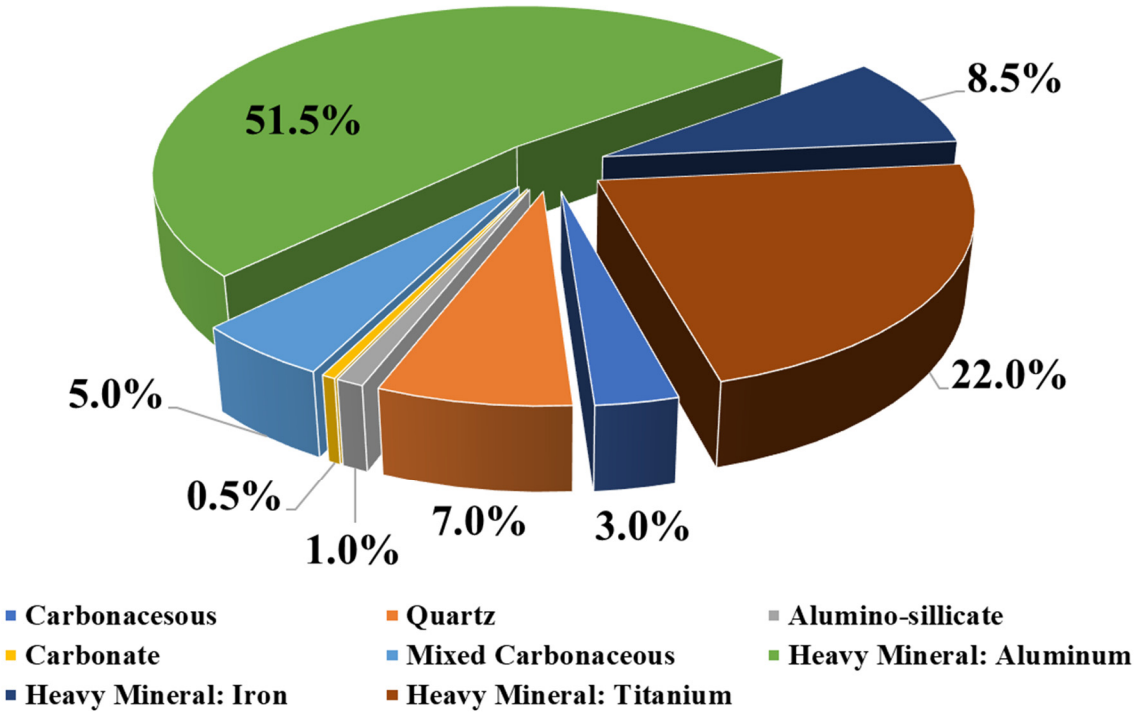


Figure 19. Error distribution of the *k*-NN ML model over the eight mineral categories.

Table 4 illustrates the original training data set and test data set that were used in model fitting. The percent of heavy mineral: aluminum, heavy mineral: titanium, and heavy mineral: iron are all less than 1%. In order to assess the influence of the data volume on the *k*-NN model performance, we calculated the accuracy from the confusion matrix.

Table 4. Original training data set and test data set.

Training Data Set

Name	Count	Percent
Alumino-silicate	29313	31.16%
Carbonaceous	27227	28.94%
Carbonate	21825	23.20%
Heavy Mineral: Aluminum	24	0.03%
Heavy Mineral: Iron	680	0.72%
Heavy Mineral: Titanium	45	0.05%
Mixed Carbonaceous	9153	9.73%
Quartz	5804	6.17%

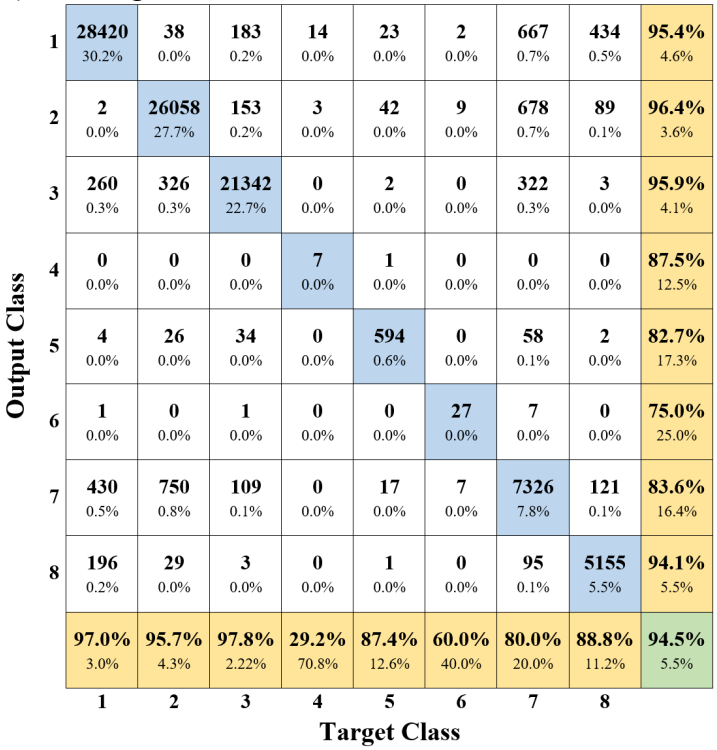
Test Data Set

Name	Count	Percent
Alumino-silicate	3000	39.64%
Carbonaceous	2600	34.36%
Carbonate	1600	21.14%
Heavy Mineral: Aluminum	2	0.03%
Heavy Mineral: Iron	4	0.05%
Heavy Mineral: Titanium	2	0.03%
Mixed Carbonaceous	260	3.44%
Quartz	100	1.32%

Figure 20 is the confusion matrixes of training data and test data. In ML, the confusion matrix is used to demonstrate the accuracy, recall, and precision of the ML model performance. Specifically, "Output class" is the class label predicted by the ML model; "Target class" is the actual class labels determined by the SEM images (i.e., the "ground truth"). The class labels from 1 through 8 are based on the definition listed in Table 4 (i.e., Class 1: alumino-silicate; Class 2: carbonaceous; Class 3: carbonate; Class 4: heavy mineral – aluminum; Class 5: heavy mineral – iron; Class 6: heavy mineral – titanium; Class 7: mixed carbonaceous; Class 8: quartz). The "recall" rate is defined as the total number of correctly classified X samples divide by the total number of X samples (here, X can be any class from Class 1 through Class 8). The "precision" rate is defined as the ratio of the total number of correctly classified X samples to the total number of predicted X samples (here, X can be any class from Class 1 through Class 8). Therefore, based on these definitions, the yellow column-vector shows the "precision" of the ML model, whereas the yellow row-vector shows the "recall" of the ML model.

Illustrated by the confusion matrixes, the overall accuracy in training data (94.5%) and test data (96.2%) are both high enough. However, the precision of some specific classes, like Aluminum, Iron, and Titanium, is quite limited (even zero). Because *k*-NN is a voting algorithm and the data sets of these classes are quite small, the weights of small sets are also small. The result confirms that limited data amount is not beneficial to the model accuracy and precision. We then enlarged the data sets of these three classes (Aluminum, Iron, and Titanium) to check if oversampling can improve the accuracy and precision of these classes.

a) Training confusion matrix



b) Test confusion matrix

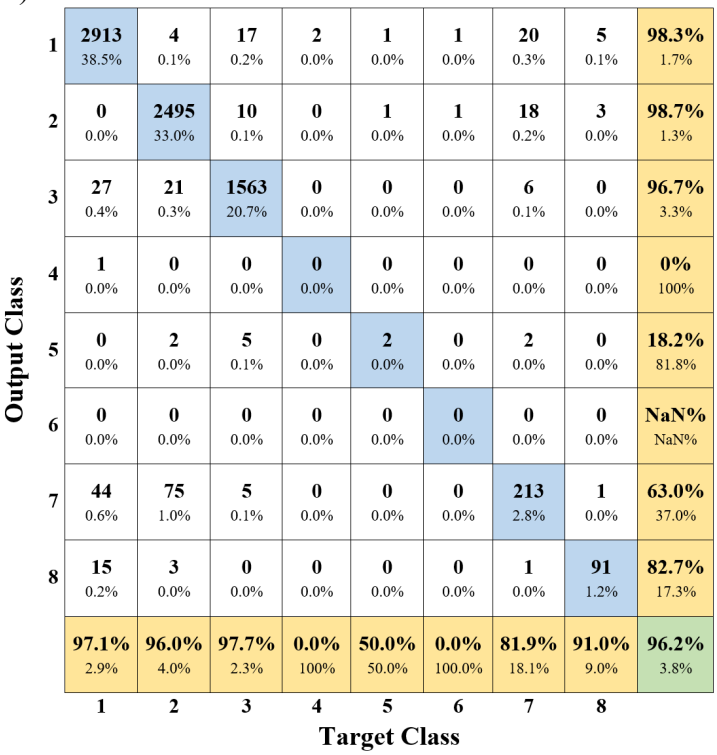


Figure 20. Training confusion matrix (a) and test confusion matrix (b) of the *k*-NN ML model with eight element features. The class labels are: 1. Alumino-silicate; 2. Carbonaceous; 3. Carbonate; 4. Heavy Mineral: Aluminum; 5. Heavy Mineral: Iron; 6. Heavy Mineral: Titanium; 7.

Mixed Carbonaceous; 8. Quartz. The green grid block shows the overall accuracy of the ML model. The yellow column-vector shows the precision of the ML model, whereas the yellow row-vector shows the recall of the ML model. In the blue and white grid blocks, the number is the microparticle sample number and the percentage shows the ratio of sample number to the total sample number.

Table 5 illustrates the enlarged training data set and enlarged test data set. The data amount of the three classes, Aluminum, Iron, and Titanium, have been enlarged. The enlarged data amount is 300 times of the original Aluminum data volume, 20 times of the original Iron data volume, and 200 times of the original Titanium data volume. After this data oversampling process, we calculated the accuracy again from the updated confusion matrix.

Table 5. Enlarged training data set and enlarged test data set.

Training Data Set

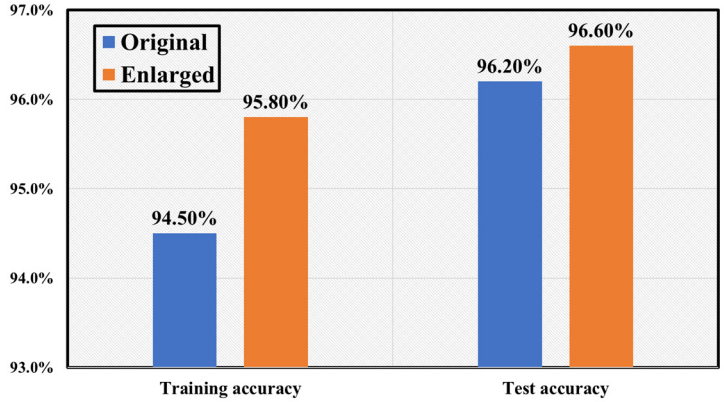
Name	Count	Percent
Alumino-silicate	29313	23.75%
Carbonaceous	27227	22.06%
Carbonate	21825	17.69%
Heavy Mineral: Aluminum	7600	6.16%
Heavy Mineral: Iron	13280	10.76%
Heavy Mineral: Titanium	9200	7.46%
Mixed Carbonaceous	9153	7.42%
Quartz	5804	4.70%

Test Data Set

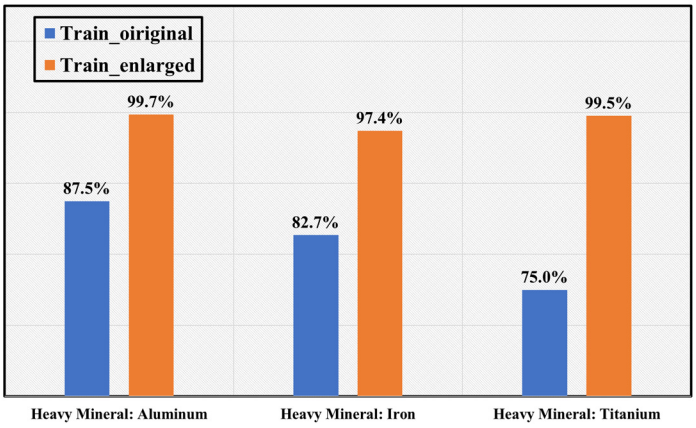
Name	Count	Percent
Alumino-silicate	3000	35.89%
Carbonaceous	2600	31.10%
Carbonate	1600	19.14%
Heavy Mineral: Aluminum	200	2.39%
Heavy Mineral: Iron	400	4.78%
Heavy Mineral: Titanium	200	2.39%
Mixed Carbonaceous	260	3.11%
Quartz	100	1.20%

Figure 21 illustrates the overall accuracies of the *k*-NN model and the specific class prediction accuracies before and after oversampling. The overall accuracy goes higher after oversampling. Moreover, the prediction precisions of the three specific classes are also improved, especially in the test data set. It confirms that enlarging the data volumes of some classes can improve the ML model performance.

a)



b)



c)

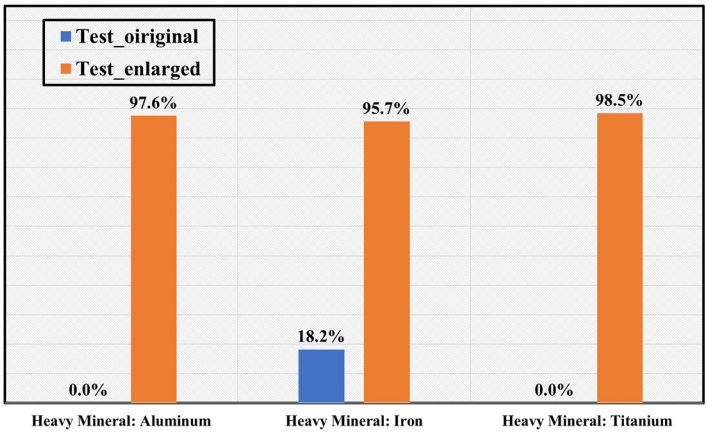


Figure 21. a) Overall accuracy of the *k*-NN ML model, and the comparison between the accuracies of the original data set and the enlarged data set in the b) training data, and c) test data.

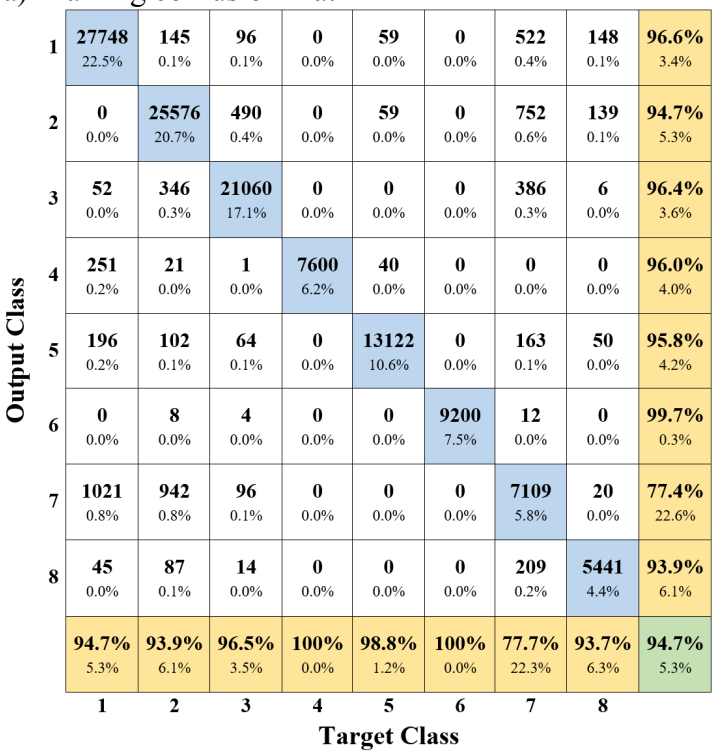
To achieve a higher prediction accuracy in the ML processes, we added more input features into the model. Studies in the literature have used features to define the geometry factor of particles, including the aspect ratio, shape factor, convexity, etc. In this project, we selected the aspect ratio, shape factor, and convexity as the microparticle geometry features. Therefore, the ML model has 11 features (i.e., model inputs) after accounting for the particle geometry features. These feastures are the eight chemical elements (O, Al, Si, C, Mg, Ca, Ti, and Fe) and the three measurement factors (aspect ratio, shape factor, and convexity). **Table 6** illustrates the training and test confusion matrices of the *k*-NN ML model using the 11 features. The overall accuracy of the ML model is lower than the model using only the 8 chemical features. Moreover, the ML model prediction precision of Mixed Carbonaceous and Quartz from the training data set is 41.6% and 14.9%, which are lower than the results of 75.0% and 44.3% from the test data set. The ML model precision is satisfying in the test data set but unsatisfying in the training data set, which implies the overfitting issue in the k-NN model. To avoid overfitting, we can enlarge data sets, reduce feature dimension, or change the ML model. However, because the data have been balanced, enlarging the data set cannot improve the precision in this situation. We also cannot reduce the data dimension because we need to have a model with total 11 features. The only approach is to change the ML model. Rather than using the *k*-NN ML model, we decided to switch to a ML model based on the SVM algorithm.

Table 6. Training and test confusion matrices of the *k*-NN ML model using the 11 features.

	Training confusion matrix	Test confusion matrix
Precision of Mixed Carbonaceous	41.6%	75.0%
Precision of Quartz	14.9%	44.3%
Overall accuracy	83.6%	85.9%

Figure 22 illustrates the training and test confusion matrices from the SVM ML model with 11 features in total. The overall accuracy of the training and test data sets from the SVM model is 94.7% and 94.5%, respectively, which are both higher than the overall accuracies of the *k*-NN model. Because most of the precisions are satisfying, we can now extract the hidden correlations between microparticle geometry and its chemical composition, which means that using the SVM model we can tell if a specific element is more likely to be associated with a particular particle size or shape.

a) Training confusion matrix



b) Test confusion matrix

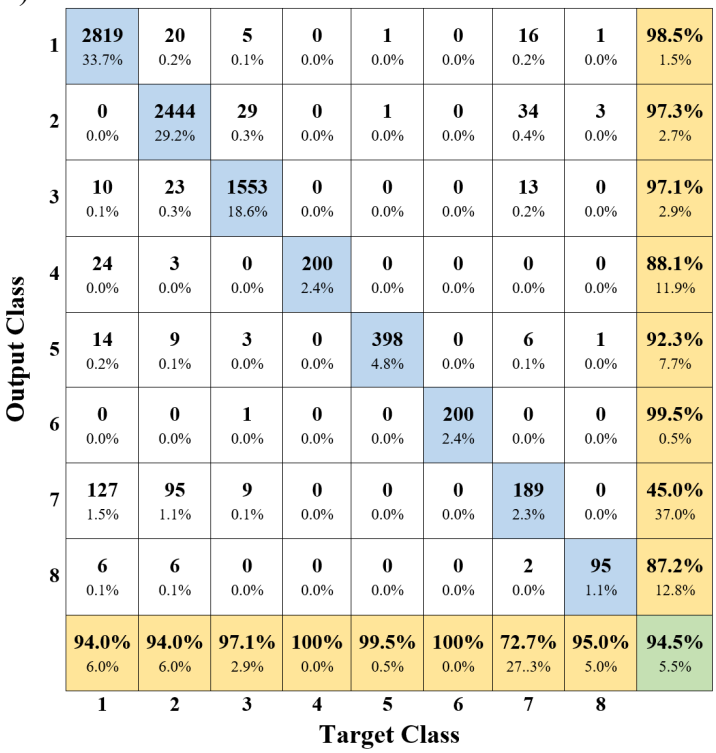


Figure 22. Training confusion matrix (a) and test confusion matrix (b) of the SVM model with total 11 features (ML model inputs). The class labels are: 1. Alumino-silicate; 2. Carbonaceous; 3. Carbonate; 4. Heavy Mineral: Aluminum; 5. Heavy Mineral: Iron; 6. Heavy Mineral: Titanium; 7.

Mixed Carbonaceous; 8. Quartz. The green grid block shows the overall accuracy of the ML model. The yellow column-vector shows the precision of the ML model, whereas the yellow row-vector shows the recall of the ML model. In the blue and white grid blocks, the number is the microparticle sample number and the percentage shows the ratio of sample number to the total sample number.

Based on the confusion matrix results shown in Figure 22. The ML model prediction precision of the Mixed Carbonaceous class is unsatisfying. Meanwhile, the details of Mixed Carbonaceous are not as important as other components in practical applications. Therefore, we merged Mixed Carbonaceous with Alumino-silicate. It should be noted that the merge of Class 7 (the class of mixed carbonaceous) into the class of alumino-silictae is for the purpose of data analytics; the ML model precision rate for the class of mixed carbonaceous is relatively low compared to the other seven classes. Due to the same reason, we also combined the three Heavy Mineral classes. The remaining classes are thus Quartz (Q), Alumino-silicate (AS), Heavy Mineral (HM), Carbonate (CB), and Carbonaceous (CBN). **Figure 23** is the binary tree which expresses the geometry-chemical correlation in a mathematical way.

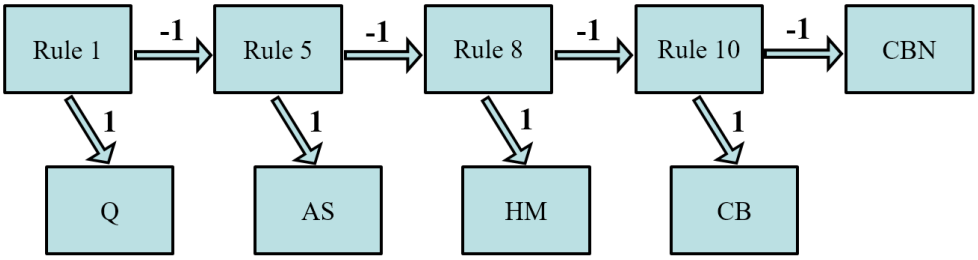


Figure 23. A binary tree which expresses the geometry-chemical correlation in a mathematical way.

Table 7 illustrates the ML-model-fitted weight coefficients in Equation 6, which is a mathemetical way to describe the geometry-chemical correlation. The data normalization process is based on the equation of $(X - \mu) / \sigma$. These coefficients, associated with Equations 5 and 6 and the workflow shown in Figure 6, can determine the class of a new, unknown data point. Specifically, eight element features and one geometry feature (the aspect ratio) were used as the ML model inputs. In this way, the geometry-chemical (geometry-label) correlation is established.

Table 7. The ML-model-fitted weight coefficients in Equation 6, which is a mathemetical way to describe the geometry-chemical correlation. There nine weigting coefficients; eight of them are for the chemical features and one of them is for the geometric feature. The geometry and chemical compostion is related in this way. These coefficients, associated with Equations 5 and 6 and the workflow shown in Figure 6, can determine the class of a new, unknown data point.

	Weights	Bias	μ	σ
Rule 1	[0.64, -13.02, 2.98, 1.05, -2.36, -0.17, -0.17, -0.00, -0.02]	-8.59	[0.25, 0.01, 0.02, 0.69, 0.00, 0.00, 0.00, 1.72, 2.47]	[0.05, 0.01 0.03, 0.11, 0.03, 0.00, 0.01, 0.46, 1.43]
Rule 2	[0.63, -0.91, 0.84, 0.54, -0.571, -1.49, -2.97, -0.04, 0.27]	-1.00	[0.22, 0.00, 0.01, 0.74, 0.00, 0.00, 0.01, 1.61, 1.95]	[0.03, 0.01, 0.02, 0.05, 0.01, 0.00, 0.01, 0.37, 0.89]
Rule 3	[0.94, -0.21, 2.19, 1.59, -5.67, -0.02, 0.03, -0.00, 0.03]	-4.24	[0.25, 0.00, 0.00, 0.70, 0.02, 0.00, 0.00, 1.61, 2.05]	[0.05, 0.00, 0.01, 0.04, 0.00, 0.00, 0.00, 0.40, 1.12]
Rule 4	[-0.07, -1.09, 4.36, -1.20, -0.28, -0.06, -0.23, -0.09, -0.01]	-2.34	[0.20, 0.00, 0.00, 0.78, 0.00, 0.00, 0.00, 1.75, 1.95]	[0.02, 0.00, 0.01, 0.03, 0.00, 0.00, 0.00, 0.52, 1.05]
Rule 5	[0.85, 1.51, -0.14, 0.18, -0.04, -0.29, -1.32, 0.05, 0.23]	1.99	[0.25, 0.01, 0.02, 0.69, 0.01, 0.00, 0.00, 1.70, 2.46]	[0.06, 0.01, 0.03, 0.11, 0.03, 0.00, 0.01, 0.45, 1.41]
Rule 6	[-1.05, 10.06, 2.25, -1.00, -0.70, 0.03, 0.10, -0.03, 0.08]	6.01	[0.26, 0.01, 0.02, 0.69, 0.02, 0.00, 0.00, 1.68, 2.38]	[0.06, 0.01, 0.03, 0.10, 0.04, 0.00, 0.00, 0.45, 1.38]
Rule 7	[-0.53, 8.74, 5.89, -2.76, 1.32, 0.11, -0.02, -0.04, -0.01]	6.50	[0.23, 0.01, 0.017, 0.73, 0.01, 0.00, 0.00, 1.75, 2.31]	[0.05, 0.01, 0.03, 0.10, 0.03, 0.00, 0.00, 0.50, 1.35]
Rule 8	[-0.56, 0.27, 0.10, -0.40, -0.52, 0.52, 1.88, 0.02, -0.18]	-0.98	[0.26, 0.00, 0.00, 0.71, 0.03, 0.00, 0.01, 1.60, 2.05]	[0.06, 0.00, 0.01, 0.08, 0.03, 0.00, 0.01, 0.37, 1.08]
Rule 9	[-0.42, 0.41, -0.18, -0.72, 0.03, 1.10, 3.97, -0.13, -0.04]	-3.04	[0.20, 0.00, 0.00, 0.78, 0.00, 0.00, 0.01, 1.73, 1.96]	[0.02, 0.00, 0.01, 0.03, 0.00, 0.00, 0.01, 0.51, 1.03]
Rule 10	[-1.35, -0.63, -0.73, -2.75, 9.65, 0.00, -0.07, -0.02, -0.12]	3.38	[0.23, 0.00, 0.00, 0.75, 0.02, 0.00, 0.00, 1.70, 2.01]	[0.05, 0.00, 0.00, 0.08, 0.03, 0.00, 0.00, 0.48, 1.11]

5.3.2. Conclusions and Summary

For Goal 3, we first utilized an unsupervised ML method, the *k*-means clustering scheme, for preliminary data examination. The unclassified particle data was divided into classes using the SV value. This unsupervised learning illustrates our capability to split unknown particles. We then used the confusion matrices to check the precision, recall, and accuracy of the training data sets and test data sets. For the data with low precision, we enlarged their data volume using the oversampling method to improve their precision and overall accuracy. To avoid overfitting, we switched the ML model from k-NN to SVM. Based on the SVM framework, we developed the geometry-chemical correlation in a mathematical way.

Specifically, in this project, we initially selected the aspect ratio, shape factor, and convexity as the microparticle geometry features. Therefore, the initial ML model has 11 input features (i.e., model inputs) after accounting for the particle geometry features. These 11 feastures are the eight chemical elements (O, Al, Si, C, Mg, Ca, Ti, and Fe) and the three measurement factors (aspect ratio, shape factor, and convexity). After ML model training, we found that the shape factor and convexity had relatively lower influence on the model. The particle size also had a lesser effect on the ML model output. Therefore, in the final ML model, only the eight element features and one geometry feature (the aspect ratio) were used as the ML model inputs, leading to total nine weight coefficients as shown in Table 7. Note that the chemical-geometry correlation is expressed in an implicit way. Specifically, these weight coefficients, associated with Equations 5 and 6 and the workflow shown in Figure 6, can determine the class of a new, unknown microparticle sample.

5.4. *Goal 4*: Develop advanced numerical modeling capabilities to improve the fundamental understanding of microparticle transport and deposition (aerodynamic properties) in the human lung.

5.4.1. Results and Discussion

Figure 24 illustrates the LB-simulated air pressure distributions within the pore space of 2D cross sections of simplified human lung models having pore diameter coefficient of variation (COV) of 5% and 25%. The pore diameter COV is defined as the ratio of the standard deviation of pore diameter to the mean pore diameter, which indicates how heterogeneous the pore size is in the lung model. A larger pore diameter COV suggests a more heterogeneous pore size distribution in the human lung model. The air pressure is presented in the LB unit. It can be observed that the human lung model having a pore diameter COV of 25% had a more heterogeneous pore size distribution, because a higher pore diameter COV leads to a wider pore diameter distribution.

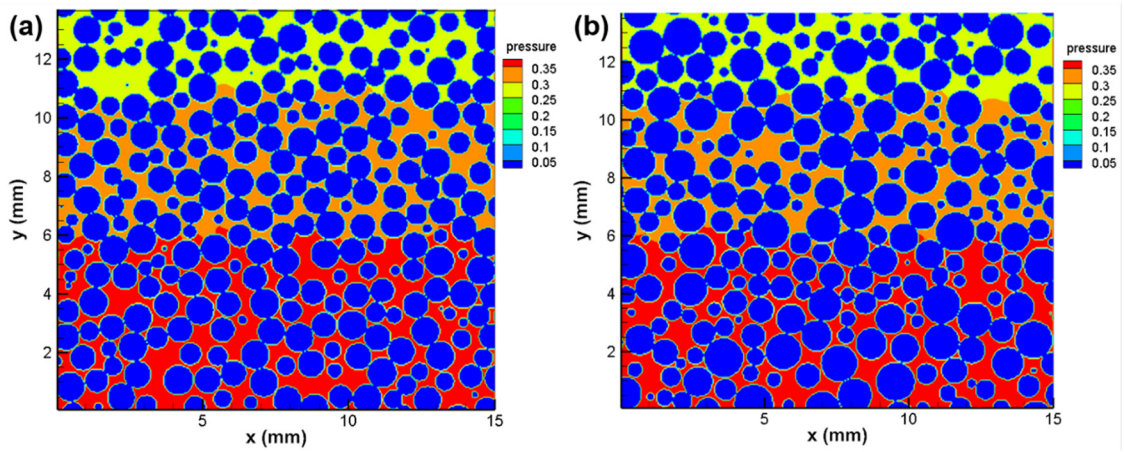


Figure 24. LB-simulated air pressure distributions within the pore spaces of simplified human lung models having a) 5% pore diameter COV, and b) 25% pore diameter COV. The pore diameter COV is defined as the ratio of the standard deviation of pore diameter to the mean pore diameter, which indicates how heterogeneous the pore size is in the human lung model. A larger pore diameter COV suggests a more heterogeneous pore size distribution in the human lung model.

Figure 25 illustrates the pore-scale LB-simulated air flow velocity magnitude distributions and the associated air streamlines within a 2D cross section of the 3D human lung model having pore diameter COV of 5%. The air flow velocity magnitude was presented in the LB unit. Figure 25 illustrates that the pore-scale air flow was well resolved and simulated by the LB model.

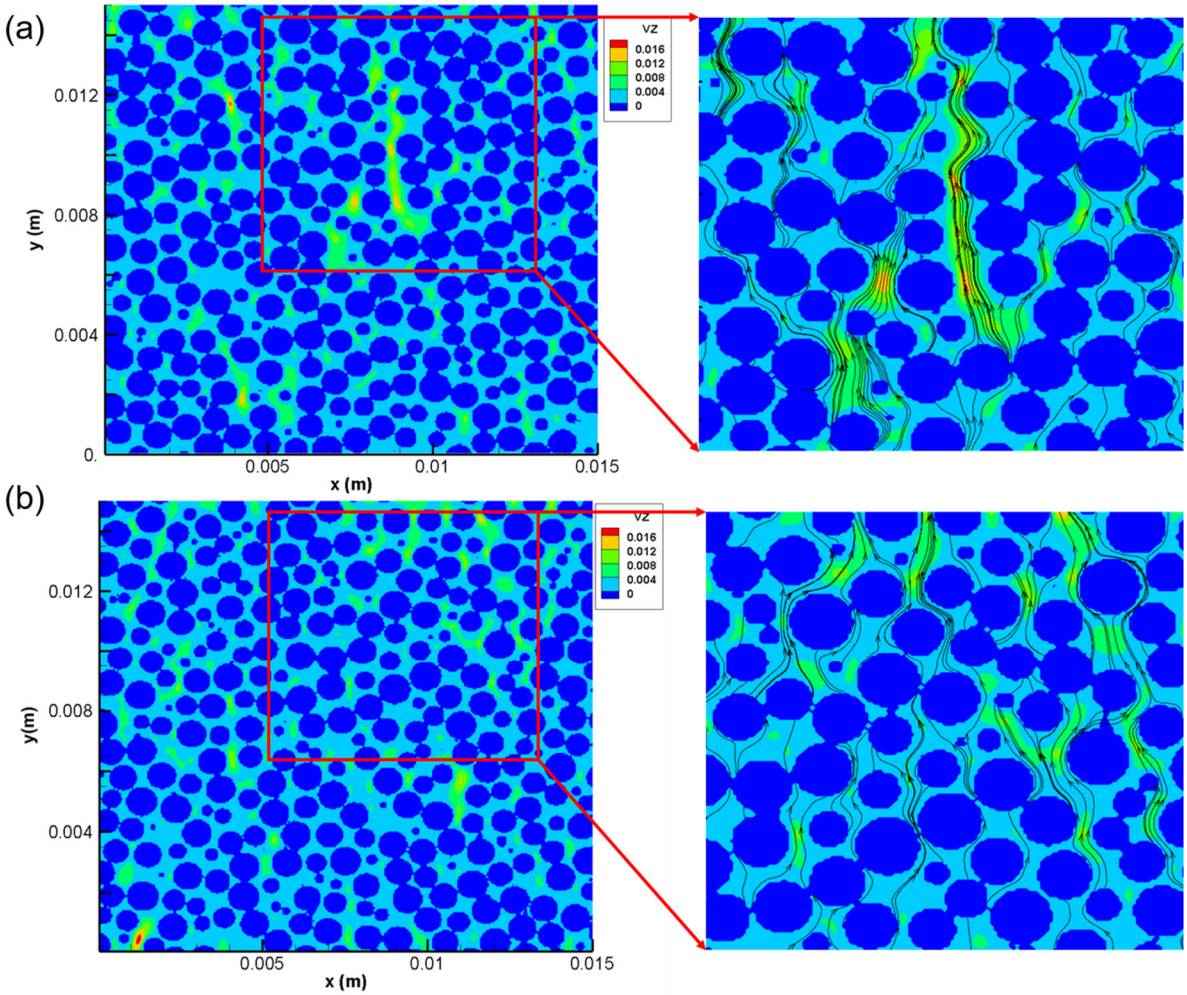


Figure 25. LB-simulated air flow velocity magnitude distributions (left panels) and the associated air streamlines (right panels) within the pore spaces of the 3D human lung model having a pore diameter COV of 5%. (a) and (b) show two regional areas of interest in this human lung model.

The Monte Carlo (MC) simulation was used to track the transport and deposition of 10,000 microparticles in the pore space of a 3D human lung model based on the LB-simulated pore air flow field and the particle tracking algorithm illustrated in Equation 18. At the end of the MC simulation, the distribution of all collision positions in the longitudinal flow (x) direction is analyzed and ordered. **Figure 26** demonstrates the cumulative distribution function (CDF) and the corresponding probability density function (PDF) of the collision positions in the x direction in the 3D human lung model. The PDF curve is the spatial derivative of the CDF curve in the main flow (x) direction. In this realization, the human lung model has an average pore diameter of 0.63 mm and a pore diameter COV of 5%. Specifically, the migration distance is the longitudinal distance between the air flow inlet and the location where the microparticle collides with the lung inner wall surfaces. The CDF indicates the probability that a microparticle travels a longitudinal distance less than x in the 3D human lung model. The PDF is the first-order derivative of the CDF curve,

which indicates the likelihood of a microparticle deposits at the longitudinal distance of x from the air flow inlet. The PDF curve is then fitted using the macroscopic particle deposition model (Equation 15) to determines the deposition coefficient, λ, in the exponential function.

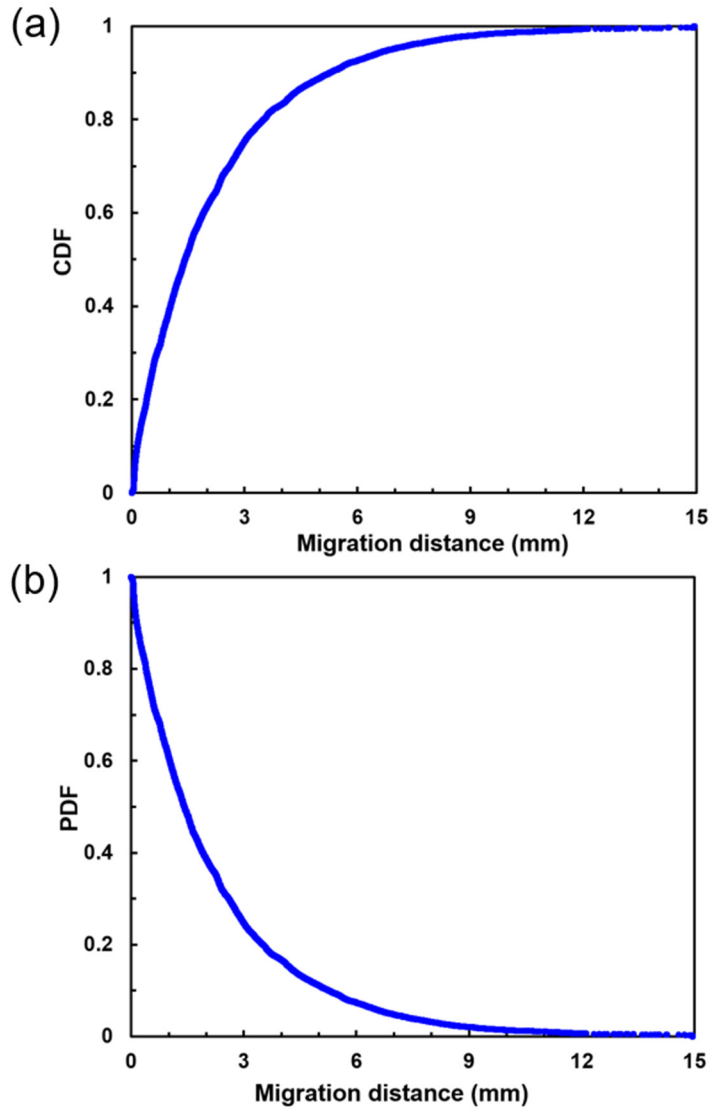


Figure 26. (a) Cumulative distribution function (CDF) and (b) probability density function (PDF) as a function of microparticle migration distance in a 3D human lung model. These two functions were obtained based on MC simulations of the transport and deposition of 100,000 microparticles. The PDF curve is the spatial derivative of the CDF curve in the main flow (x) direction. The PDF curve can be fitted using the macroscopic particle deposition model (Equation 15) to determines the deposition coefficient, λ, in the exponential function. In this realization, the human lung model has an average pore diameter of 0.63 mm and a pore diameter COV of 5%.

Figure 27 presents the deposition coefficient, obtained by fitting the pore-scale-simulated microparticle migration distances, as a function of microparticle diameter in human lung models having two pore diameter COVs (5% and 25%). The same air flow rate boundary condition was imposed on both lung models having different pore size COVs. Non-monotonic evolution of the deposition coefficient as a function of microparticle diameter is observed, which leads to the classic "U" shape curve. Compared to intermediate-sized particles, the ultra-fine particles with diameter smaller than 0.01 μm have higher deposition coefficients because Brownian motion dominates their deposition. In addition, large particles with diameter bigger than 1 μm also have higher deposition coefficients than the intermediate-sized particles because gravitational settling dominates the deposition of large particles.

Figure 27 also illustrates that the deposition coefficients in the lung model having pore diameter COV of 5% are larger than those in the lung model with pore diameter COV of 25% for all microparticle sizes. This is because the human lung model with pore diameter COV of 25% has more large-sized pores because of the uniform distribution of the pores. As a consequence, relatively large flow channels were formed in the lung model having a 25% pore diameter COV, leading to more preferential flow paths than the lung model having a 5% pore diameter COV, which is favorable for microparticle transport through the pore spaces and thus results in a smaller deposition coefficient in the human lung model having a 25% pore diameter COV.

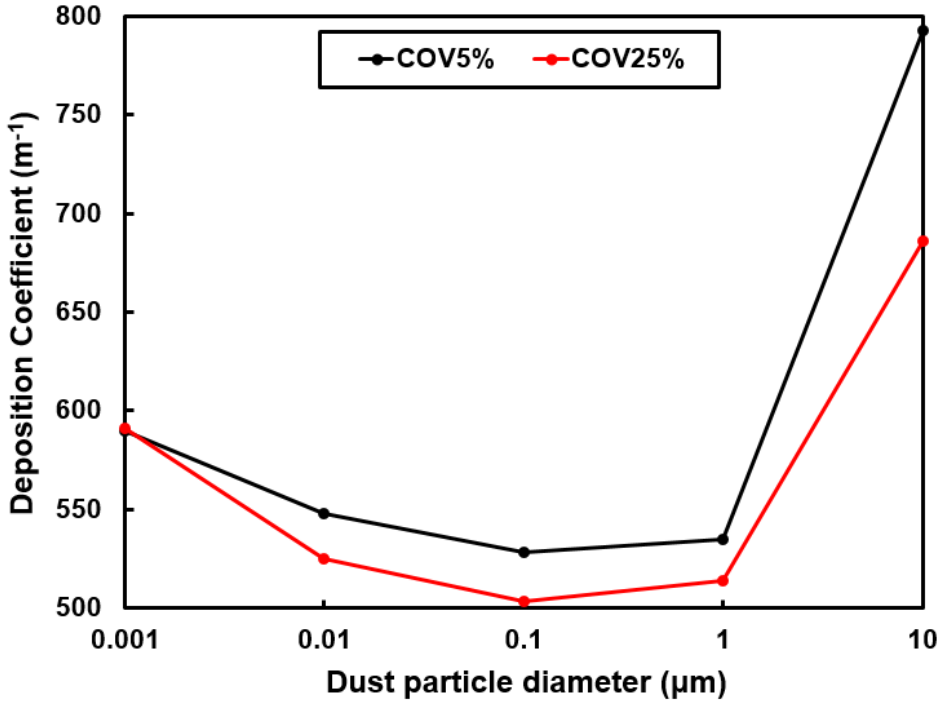


Figure 27. Deposition coefficient as a function of microparticle diameter in human lung models having pore diameter COVs of 5% and 25%. The same air flow rate was imposed on both the lung models as the boundary condition. These two curves were simulated using the developed LB air flow model and the microparticle tracking algorithm in 3D human lung models. The "U-shape" of the curves is caused by the enhanced deposition rate of ultrafine particles due to Brownian motion and enhanced deposition rate of relatively large particles due to gravitational settling.

Figure 28 shows that various whole-lung models have demonstrated the "U-shape" curve for the prediction of dust deposition amount as a function of the microparticle diameter ranging from 1 nm to 10 μm. The typical U-shape curve results from the fact that microparticles smaller than 0.1 μm are dominated by Brownian motion and particles larger than 1 μm are dominated by gravitational settling; both mechanisms enhance the total deposition of microparticles in the human lung. It is clear that our fundamental microparticle transport prediction shown in Figure 27, which is based on pore-scale LB modeling of air flow and microparticle tracking, is consistent with the U-shape curves predicted by other lung models illustrated in Figure 28.

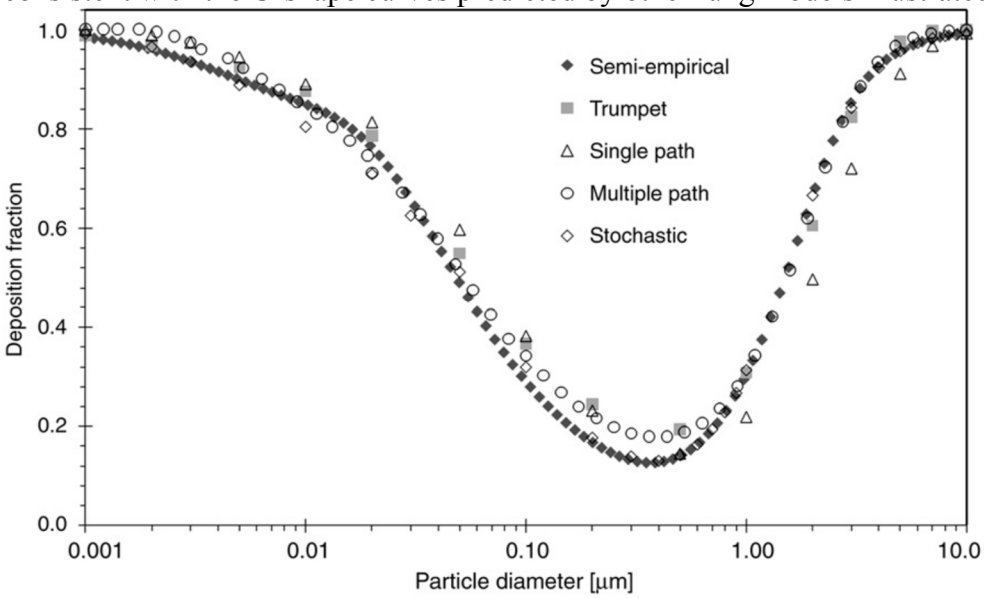


Figure 28. Whole-lung model predictions of dust deposition amount as a function of microparticle diameter. Five models are presented: semi-empirical (ICRP, 1994), trumpet (Choi and Kim, 2007), single path (Hofmann, 1982), multiple path (Asgharian et al., 2001), and stochastic (Koblinger and Hofmann, 1990). Figure from Hofmann (2011). The "U-shape" curve is caused by the enhanced deposition rate of ultrafine particles due to Brownian motion and enhanced deposition rate of relatively large particles due to gravitational settling.

6. Technology Readiness Assessment

The specific research goals in this project are:

1) Develop 3D non-destructive, element-specific CT capabilities to identify element heterogeneity in a single mine dust particle at the spatial resolution of 50 nm/pixel.
2) Develop autonomous image pattern recognition capability to extract information for dust particles, such as size, shape, density, and element.
3) Identify potentially predictive correlations between particle size, shape, and chemical composition using machine learning and big data analytics.
4) Develop advanced numerical modeling capabilities to improve the understanding of mine dust transport and deposition (aerodynamic properties) in the human lung.

For the first goal, we developed an AI-based imaging and segmentation technology to evaluate element distribution within the 3D structural space of a microparticle. In this technology, Nano-CT and SEM are used to scan the same area of microparticle surface, in order to collect training data that contain the correlations between the greyscale CT values and the SEM element information. In the ML training process, not only the contrast in greyscale CT values is used as a training feature, but also the geometrical characteristics of the interfaces between elements are extracted for training. Next, a random decision forest training and classification process is performed to segment the greyscale CT pixels throughout the entire 3D structural space within the microparticle. In this study, we use 200 decision trees in the random decision forest model, and the final decision is made by voting. The AI model was used to segment element distribution in two custom-made microparticles, and the evaluated surface coating thicknesses were in good agreement with the values provided by the manufacturer. Therefore, we think that the AI-based element segmentation method is promising and works well in some specific scenarios. However, more fundamental research is needed along this direction. In order to make this AI-based method to work, in the ML training process, not only the contrast in grayscale CT values is used as a training feature but also the geometrical characteristics of the interfaces between minerals are extracted for training. This suggests that we will need at least two or three minerals present at the microparticle surface so that the SEM scanning can see them and label them as the ML features (i.e., the ML model inputs). These features not only include the contrast in grayscale CT values but also account for the geometrical characteristics of the interfaces between minerals. The technology readiness of the autonomous image processing software (Goal 2) is dependent on the improvement of the AI-based mineral segmentation capability (Goal 1).

The other two goals (Goals 3 and 4) have relatively higher technology readiness and can be applied to various scenarios as long as the data sets are available. Specifically, for the third goal, various ML models have been tested for classification of dust particles. The method based on k-means, k-NN, and SVM have been developed and the classification results are satisfying by comparisons with labeled microparticle data sets. We found that the SVM method provides an overall training and testing accuracy about 10% higher than the k-NN, because the SVM mitigates the overfitting issue better. In addition, the SVM model accounts for the geometric property of particles, which implies that there are underlying correlations between particle geometry and chemical composition. For the fourth goal, we conducted fundamental fluid dynamics and particle transport simulations at the pore scale in a synthesized human lung model. The air flow was simulated using the LB method, which is a numerical model for solving air flow at the pore scale. Dust particle migration in the human lung model was then simulated using the particle tracking method based on the LB-simulated air flow field. Three particle transport

and filtration mechanisms were accounted for in particle tracking, including Brownian motion, streamline advection, and gravitational settling. The simulated dust particle deposition amount was plotted as a function of dust particle size, and the plot showed a “U” shape, which is consistent with the classic theoretical prediction. In these LB simulations, we generated synthesized human lung models having varying pore size distributions to study their influence on the dust particle deposition amount. The developed LB numerical air flow model and microparticle tracking numerical model will have direct applications to study dust particle transport and deposition at the pore to regional scales.

Also, it should be noted that, although the ML model has been developed, new microparticle data can still be input into the ML model to update the model weight coefficients to reflect the properties of new samples. In addition, after the testing of the two well-controlled custom-made microparticles, we also used SEM, Nano-CT, and the AI model to analyze real dust particles. However, it turned out that these real particles happened to be pure silica particles, which did not give us the desired “heterogeneous” 3D element distribution pattern. This partially justified our decision to start the project using well-controlled, custom-made microparticles because they are able to provide heterogeneous element distribution within the 3D structure of the particle.

7. Appendices

Please see attached the support letter in PDF format provided by Brian Gobrogge, Chief Operating Officer of Cospheric LLC.

Cospheric
Innovations in Microtechnology
PO Box 6762, Santa Barbara, CA 93160
Phone: (805) 687-3747
Fax: (866) 708-0375
www.cospheric.com

Virginia Tech
Dr. Cheng Chen
100 Holden Hall, Mining & Minerals Eng,
445 Old Turner St
Blacksburg, VA 24061, United States

2019-Oct-01

The purpose of this letter is to support the measurement accuracy of the artificial intelligence (AI) based image segmentation software developed by Dr. Cheng Chen's group.
Dr. Chen ordered two types of metal coated microspheres from Cospheric LLC on 03/22/2018 listed below.

Item	Description	Estimated Coating Thickness
Custom Aluminum Coated BTGMS - 60g	Aluminum Coated BTGMS 30-100um Microspheres with 1-2% by weight of Aluminum (~400nm). Lot# 180204-1064	375nm
SLGMS-AG-2.71 45-53um - 5g	Silver Coated Solid Soda Lime Glass Microspheres 2.71g/cc 45-53um - 5g Lot# 161021-200 or 161025-300	435nm

The surface coating thickness was calculated by analyzing the difference in true particle density before and after surface coating. We use a helium gas pycnometer which measures all of the microparticle volume that is impenetrable by helium, and measure the total microparticle mass on an ultra-precision balance; the mass and volume information is then used to calculate the true particle density, assuming all particles are the estimated mean diameter.

Sincerely,
Brian Gobrogge
Brian Gobrogge

8. Acknowledgement/Disclaimer

This study was sponsored by the Alpha Foundation for the Improvement of Mine Safety and Health, Inc. (ALPHA FOUNDATION). The views, opinions and recommendations expressed herein are solely those of the authors and do not imply any endorsement by the ALPHA FOUNDATION, its Directors and staff.

References

Altman, N.S. (1992), An Introduction to Kernel and Nearest Neighbor Nonparametric Regression. The American Statistician, 46, 175-185.

Asgharian, B.,Hofmann,W., and Bergmann,R. (2001), Particle deposition in a multiple path model of the human lung. Aerosol Science and Technology, 34, 332-339.

Castranova, Vincent, and Val Vallyathan (2000), Silicosis and coal workers' pneumoconiosis, Environmental health perspectives, 108, Suppl 4: 675.

Centers for Disease Control (CDC), (2006), Advanced Cases of Coal Workers' Pneumoconiosis-Two Counties, Virginia; Report No. 55(33); MMWR: Atlanta, GA, USA; pp. 909–913.

Chawla, Nitesh V., et al. (2002), SMOTE: synthetic minority over-sampling technique, Journal of artificial intelligence research 16: 321-357.

Chen, C., Packman, A.I., and Gaillard, J.F. (2008), Pore-scale analysis of permeability reduction resulting from colloid deposition, Geophysical Research Letters, 35, L07404, doi:10.1029/2007GL033077.

Chen, C., Lau, B.L.T., Gaillard, J.F., and Packman, A.I. (2009), Temporal evolution of pore geometry, fluid flow, and solute transport resulting from colloid deposition, Water Resources Research, 45, W06416, doi:10.1029/2008WR007252.

Chen, C., and Zhang, D. (2009), Lattice Boltzmann simulation of the rise and dissolution of two-dimensional immiscible droplets, Physics of Fluids, 21, 103301, doi: 10.1063/1.3253385.

Chen, C., Packman, A.I., Zhang, D., and Gaillard, J.F. (2010), A multi-scale investigation of interfacial transport, pore fluid flow, and fine particle deposition in a sediment bed, Water Resources Research, 46, W11560, doi:10.1029/2009WR009018.

Chen, C., Zeng, L., and Shi, L. (2013), Continuum-scale convective mixing in geological CO2 sequestration in anisotropic and heterogeneous saline aquifers, Advances in Water Resources, 53, 175-187.

Chen, C. (2016), Multiscale imaging, modeling, and principal component analysis of gas transport in shale reservoirs, Fuel, 182, page 761-770, DOI: 10.1016/j.fuel.2016.06.020.

Chen, C., Z. Wang, D. Majeti, N. Vrvilo, T. Warburton, V. Sarkar, and G. Li (2016), Optimization of lattice Boltzmann simulation with Graphics-Processing-Unit parallel computing and the application in reservoir characterization, SPE Journal, 21(4), 1425-1435, SPE-179733-PA. http://dx.doi.org/10.2118/179733-PA.

Chen, S., and Doolen, G. D. (1998), Lattice Boltzmann Method for Fluid Flows, Annu. Rev. Fluid Mech., 30, 329-364.

Choi, J.-I., & Kim,C.S. (2007). Mathematical analysis of particle deposition in human lungs: An improved single path transport model. Inhalation Toxicology, 19, 925–939.

Cortes, Corinna, and Vladimir Vapnik (1995), Support-vector networks, Machine learning 20.3: 273-297.

Cospheric, www.cospheric.com.

Fan, M., et al. (2018), Interaction between proppant compaction and single-/multiphase flows in a hydraulic fracture. SPE Journal, 23(04): p. 1,290-1,303.

Ginzburg, I. (2008), Consistent lattice Boltzmann schemes for the Brinkman model of porous flow and infinite Chapman-Enskog expansion. Physical Review E, 77(6): p. 066704.

Ginzburg, I., D. d'Humières, and A. Kuzmin (2010), Optimal stability of advection-diffusion lattice Boltzmann models with two relaxation times for positive/negative equilibrium. Journal of Statistical Physics, 139(6): p. 1090-1143.

Harrison, J. C., P.S. Brower, M. D. Attfield, C. B. Doak, M. J. Keane, R. L. Grayson, and W.E. Wallace (1997), Surface composition of respirable silica particles in a set of US anthracite and bituminous coal mine dusts. Journal of aerosol science, 28(4): 689-696.

Hartigan, J.A. and Wong, M.A. (1979), Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28, 100-108. http://dx.doi.org/10.2307/2346830

He, Haibo, et al. (2008), ADASYN: Adaptive synthetic sampling approach for imbalanced learning, Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on. IEEE, 2008.

Hofmann W. (2011), Modelling inhaled particle deposition in the human lung - a review. J. Aerosol Sci.; 42: 693–724.

International Agency for Research on Cancer (IARC) (1997), IARC Monographs on the Evaluation of Carcinogenic Risks to Humans: Silica, Some Silicates, Coal Dust and Para-Aramid Fibrils; IARC Press: Lyon, France; Volume 68.

International Commission on Radiological Protection (ICRP) (1994), Human Respiratory Tract Model for Radiological Protection. ICRP Publication 66, Annals of ICRP 24, Nos.1 3. Oxford: Pergamon Press.

International Organization for Standardization (ISO) (1995), Air Quality-Particle Size Fraction Definitions for Health Related Sampling; ISO Standard 7708; ISO: Geneva, Switzerland.

Itasca Consulting Group (2008), PFC3D – Particle Flow Code in 3 Dimensions, Version 4.0 User's Manual. Minneapolis: Itasca.

Johann-Essex, V.; Keles, C.; Sarver, E. (2017), A Computer-Controlled SEM-EDX Routine for Characterizing Respirable Coal Mine Dust. Minerals, 7, 15.

Koblinger, L., and Hofmann,W. (1990), Monte Carlo modelling of aerosol deposition in human lungs. Part I: Simulation of particle transport in a stochastic lung structure. Journal of Aerosol Science, 21, 661–674.

Kohavi, R. (1995), A study of cross-validation and bootstrap for accuracy estimation and model selection, International Joint Conferences on Artificial Intelligence, Vol. 14. No. 2.

Laney, A. and Attfield, M. (2014), Examination of potential sources of bias in the US coal workers' health surveillance program. Am. J. Public Health, 104, 165–170.

Laney, A.; Wolfe, A.; Petsonk, E.; Halldin, C. (2012), Pneumoconiosis and Advanced Occupational Lung Disease among Surface Coal Miners—16 states, 2012–2011; Report No. 61(23); MMWR: Atlanta, GA, USA; pp. 431–434.

Li, Y. and P. Huang (2008), A coupled lattice Boltzmann model for advection and anisotropic dispersion problem in shallow water. Advances in Water Resources, 31(12): p. 1719-1730.

McDowell-Boyer, L.M., J.R. Hunt, and N. Sitar (1986), Particle transport through porous media. Water Resources Research, 22(13): p. 1901-1921.

Occupational Safety and Health Administration (OSHA) (2010), Occupational Exposure to Respirable Crystalline Silica—Review of Health Effects Literature and Preliminary Quantitative Risk Assessment; OSHA:Washington, DC, USA.

Picard, R. R., and K. N. Berk (1990), Data splitting, The American Statistician 44.2: 140-147.

Pollock, D.; Potts, J.; Joy, G (2010), Investigation into dust exposures and mining practices in mines in the Southern Appalachian Region. Min. Eng., 62, 44–49.

Rousseeuw, Peter J. (1987), Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of computational and applied mathematics 20: 53-65.

Sellaro, R.; Sarver, E.; Baxter, D. (2015), A Standard Characterization Methodology for Respirable Coal Mine Dust Using SEM-EDX. Resources, 4, 939-957.

Suarthana, E.; Laney, A.; Storey, E.; Hale, J.; Attfield, M. (2011), Coal workers' pneumoconiosis in the United States: Regional differences 40 years after implementation of the 1969 Federal Coal Mine Health and Safety Act. Occup. Environ. Med., 68, 908–913.

Succi, S. (2001), The Lattice Boltzmann Equation for Fluid Dynamics and Beyond, Oxford Univ. Press, New York.

Succi, S., Benzi, R. and Higuera, F. (1991), The Lattice-Boltzmann Equation—a New Tool for Computational Fluid Dynamics. Physica D 47:219–30.

World Health Organization (WHO) (1999), Hazard Prevention and Control in the Work Environment: Airborne Dust; Report No. WHO/SDE/OEH/99.14; World Health Organization (WHO): Geneva, Switzerland; pp. 1–246.

Yao, K., M. T. Habibian, and C. R. O'Melia (1971), Water and wastewater filtration: concepts and applications, Environ. Sci. Technol., 5(11), 1105-1112.