Sabancı University Program for Undergraduate Research (PURE) Summer 2024-2025

**PROJECT/ARTICLE TITLE**

**Bushra Riaz** bushrariaz32@gmail.com *Computer Engineering, Year 3*

**Zeynep Karakaş** [zeynep.karakas@sabanciuniv.edu](mailto:zeynep.karakas@sabanciuniv.edu) *Undeclared,* Year 1

**Ayush Raj** [arindiapersonal@gmail.com](mailto:arindiapersonal@gmail.com) Electrical and Electronics Engineering. Year 4

**First - Last Name** e-mail *Program/Department, Year (For SU students: if declaration not made yet, Undeclared)*

**Dr. Melike Nur Önder**

*Materials science and nano engineering*

**Prof. Dr. Mehmet Ali Gülgün**

*Materials science and nano engineering*

**Abstract**

This study presents a machine learning based methodology for simulating lunar regolith using Earth-based geological oxide mixtures. By using X-ray diffraction (XRD) and energy-dispersive X-ray spectroscopy (EDX), dataset was prerpared. By employing k-nearest neighbors (KNN) imputation algorithm the dataset was cleaned. Principal component analysis (PCA), on original and cleaned dataset was applied to reduce dimensionality. Future research includes optimization methods that provide accurate ratios for creation of simulant using materials available in Turkey.These simulants provide a cost-effective and reliable approach for testing Moon-related activities such as construction, agriculture, and oxygen extraction on Earth.

**Keywords:** lunar regolith, machine learning, amorphous, PCA

**1 Introduction**

Lunar regolith is a layer of the Moon’s surface which is loose, heterogeneous and fragmented. It is a critical component for future space exploration and in-situ resource utilization (ISRU) (Jancsik et al., 2025).The Moon’s surface is mainly characterized by two regions: the highland and the mare. The highland is lighter in color and composed mostly of anorthositic rocks. The mare is darker, mainly composed of basalt formed by ancient volcanic activity. They originated due to meteorite impacts that created vast basins, which were later filled with lava that solidified over time (Zhang et al., 2015).

Due to the limited quantity and high value of returned Apollo and Chang’e samples, researchers rely on lunar regolith simulants to conduct experiments on Earth (Wang, 2025). These simulants are essential for testing construction techniques, agricultural potential, and oxygen extraction processes.

Previous studies have primarily focused on replicating the crystalline phases of lunar regolith using Earth-based minerals and oxides, using techniques like X-ray diffraction (XRD) and energy-dispersive X-ray spectroscopy (EDX) Zhang, He, Lin, & Hao, 2016; Yan et al., 2015).The focus of research is to determine the crystalline and amorphous components of lunar regolith. Amorphous component formed through glassy impact melts and space weathering  (Morris, 1978).It is harder to detect amorphous phase with XRD compared to crystalline phases that can be accuratly determined. For this reason amorphous phase has often been underrepresented in existing simulants, resulting in creation of simulants which lack accuracy in replicating mechanical, thermal and chemical proporties  (Blewett et al., 2021).

To address this gap, our work introduces a machine learning driven approach to accurately match both crystalline and amorphous phases of lunar regolith. We collected and processed XRD and EDX datasets from various lunar missions and Earth samples. Missing values were handled using KNN imputation (Kaushik, 2024). Two types of dataset were preapred for further use in next steps: raw and cleaned dataset. Dimensionality reduction was achieved through principal component analysis (PCA), which preserved over 90% of the dataset’s variance while improving computational efficiency.

The next step is optimization step in which convex combination formula and the Euclidean distance is used to find oxide mixture ratios. By bridging the crystalline–amorphous gap, our method advances the development of high-accuracy regolith simulants for future lunar missions.

**2 Methodology**

First, we prepared our dataset and then applied a machine learning algorithm to it. Next, we reduced the dimensionality of the dataset using PCA. Finally, we used an optimization algorithm to determine the appropriate ratios of Earth materials for producing the simulant.

**2.1 Dataset Collection**

The collected data came from techniques such as XRD and EDX, sourced from reputable literature, including Q1-level journal articles like *Advances in Space Research* and *Planetary and Space Science*, as well as NASA’s website, to ensure accuracy. Figure 1 shows part of the chemical composition daatset. For lunar and Earth samples, the lunar data included results from Apollo missions (Apollo 11, 12, 14, 15, 16, and 17), Chang’e missions (Chang’e 5 and 6), and Luna missions (Luna 16, 20, and 24). Two datasets were prepared. The first contains chemical formulas, including crystalline compounds such as SiO₂, TiO₂, and BaO, as well as amorphous element data like Cr, Co, Ni, and many more. This dataset has 100 columns and 253 rows. The second dataset focuses on mineral composition, including minerals like plagioclase and olivine, along with a column indicating the crystalline-to-amorphous ratio. This dataset has 42 columns and 161 rows. Both datasets are attached at the end of this report. Figure 3 summarizes the data collection process. We categorized the samples as lunar or Earth, and for the lunar samples, we further separated them into highland and mare regions.



Figure 1 chemical composition dataset

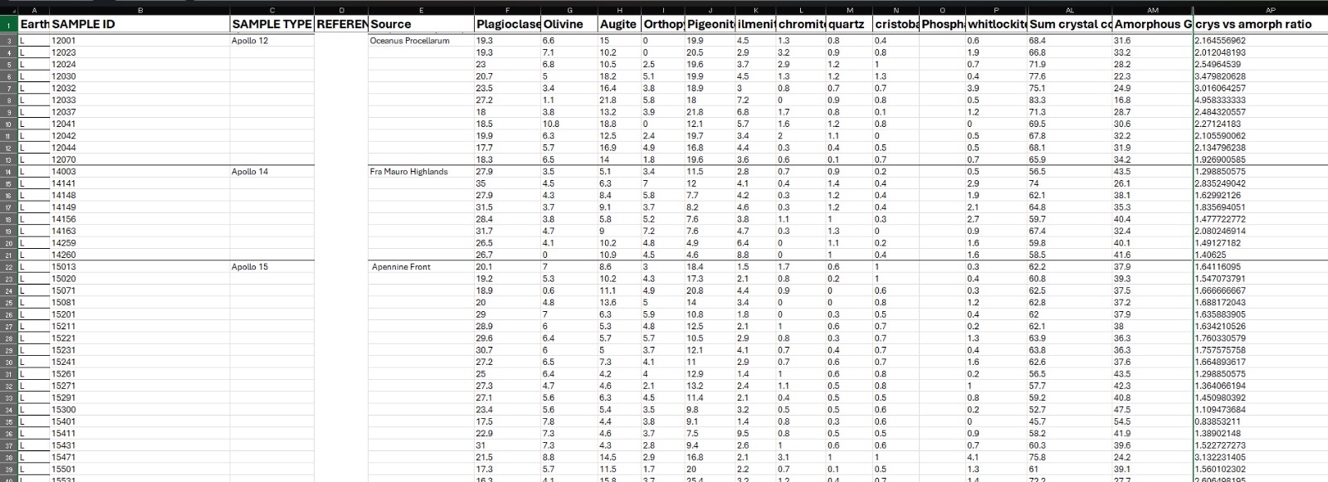


Figure 2 mineral composition dataset

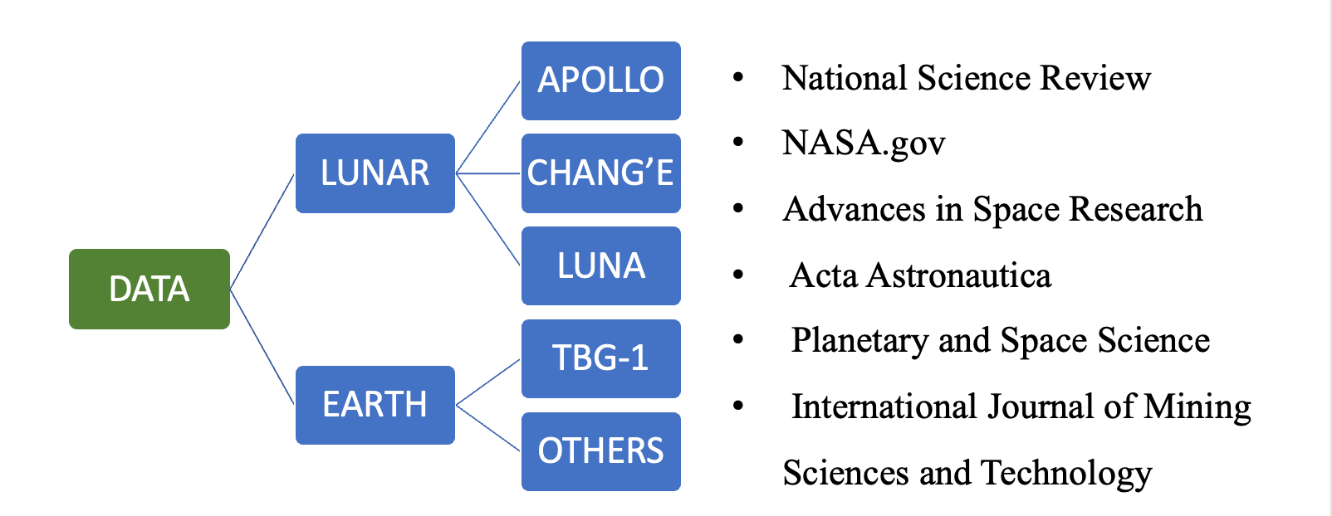


Figure 3 summary of data collection

**2.2 Process the dataset**

One of the goals of our project was to determine to what extent machine learning can help analyze the data and approximate the ratios for Earth materials. For this, we used not only the raw dataset but also a cleaned dataset created using a machine learning algorithm to fill in the missing values. We chose the k-nearest neighbors (KNN) imputation algorithm because it works efficiently on small to medium datasets and performs well when data is scarce (Troyanskaya et al., 2001). Furthermore, our dataset was entirely numerical, which is ideal for KNN imputation.

KNN imputation does not rely on correlations between features; instead, it lets the data “speak for itself.” This was important because our dataset did not exhibit strong correlations between features. KNN imputation works by storing the entire dataset, selecting a missing data point, and then finding the *k* nearest neighbors to that point based on Euclidean distance. Euclidean distance measures the straight-line distance between two points in a multi-dimensional space:

=

In this project, we used *k = 4* because it gave the highest accuracy for our predicted data values. Before applying KNN imputation, we dropped columns that had more than 40% missing values to improve accuracy. Figures 4 shows part of the KNN imputation code for both the chemical and mineral composition datasets. We applied this algorithm separately to lunar and Earth samples, and for lunar samples, we further divided them into highland and mare regions.

Other machine learning algorithms used to handle missing data include simple imputation, iterative imputation, and the random forest method (van Buuren & Groothuis-Oudshoorn, 2011; Stekhoven & Bühlmann, 2012). For simple imputation, we used methods such as mean, median, and mode. However, since our dataset had few repeating values, the mode was not an efficient choice. Furthermore, there was no distribution resembling a normal distribution in our dataset, so mean and median values were also less useful (Kamble & Deshmukh, 2017 ).Iterative imputation and random forest methods rely on correlations (linear and nonlinear) within the dataset, and their complexity is higher since they process columns iteratively. For computational efficiency, we chose KNN.



Figure 4 KNN imputation code

To evaluate the accuracy of our KNN imputation, we introduced missing values into the dataset by randomly masking a small fraction of existing non-missing entries. Since we already knew the true values for these masked points, we could directly compare the imputed results to the original values. We used Mean Squared Error (MSE) as the evaluation metric. MSE measures the average squared difference between the original values and the imputed values. A smaller MSE indicates that the imputation results are closer to the true values, meaning higher accuracy.

Mathematically:

MSE=

where is the original value and is the imputed value.

We also computed (coefficient of determination) to measure how well the imputed values explain the variation in the original data .An closer to 1 indicates a better fit. Figure 5 shows the evaluation results of our predicted values for the mineral composition dataset. KNN imputation accuracy is commonly evaluated with MSE. Zainal Abidin, Ismail, and Emran (2018), explain this in further detail in the article ‘Performance analysis of machine learning algorithms for missing value imputation’.

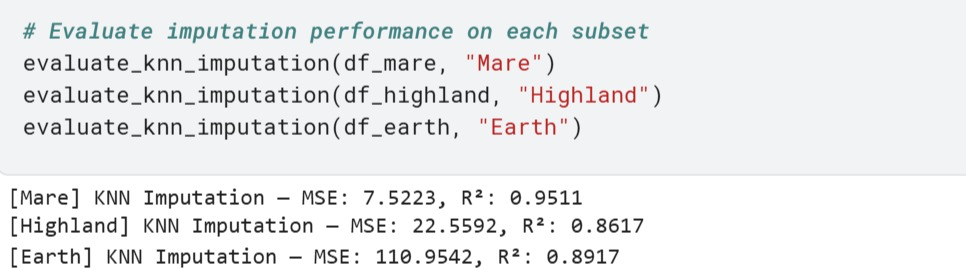


Figure 5 KNN imputation evaluation results

**2.4 Dimensionality Reduction via PCA**

To reduce the high dimensionality of the dataset while preserving its most important patterns, **Principal Component Analysis (PCA)** was applied to the cleaned and imputed data. The dataset originally contained a large number of compositional features (oxide and elemental concentrations), many of which were correlated. PCA transformed this correlated feature space into a new set of orthogonal components, called principal components (PCs), that capture the maximum variance in the data in decreasing order of importance.

The analysis showed that the first three principal components were sufficient to explain approximately **90% of the total variance** in the dataset. Specifically, **PC1** accounted for about 45% of the variance and mainly represented broad compositional differences between samples. **PC2**, contributing around 30%, captured variations that distinguished sub-groups such as Mare and Highland lunar samples. **PC3**, contributing about 15%, represented finer compositional differences.

To interpret the PCA results, both **2D and 3D visualizations** were generated. The 2D scatter plots plotted each principal component against the others (e.g., PC1 vs PC2, PC1 vs PC3) to examine clustering and overlaps between lunar and Earth samples. A **3D scatter plot** (PC1, PC2, PC3) provided an enhanced visual perspective on the separation and similarity patterns across sample types. These visual analyses helped in identifying which Earth samples were closest in composition to lunar samples, forming the basis for the subsequent optimization stage.

**2.5 Optimization for Composition Matching**

After the dimensionality reduction, an optimization algorithm was developed to determine the optimal combination of Earth-based material samples that best replicate the composition of lunar samples in the PCA space. The approach was inspired by **linear mixture modeling**, where the lunar sample’s principal component coordinates were expressed as a weighted sum of the Earth sample coordinates.

Mathematically, for a given lunar sample with PCA coordinates

These weights represent the fractional contribution of each Earth sample to the mixture, subject to the constraint that all weights are non-negative and ideally sum to 1. The problem was formulated as a **bounded least-squares optimization** and solved using the lsq\_linear solver from SciPy, ensuring that unrealistic negative proportions were avoided.

This optimization was performed for multiple lunar samples, generating a table of Earth sample weights for each case. The results revealed which Earth materials could be combined, and in what proportions, to closely match the lunar composition in the reduced PCA space. The methodology not only allowed for quantitative matching but also served as a tool for evaluating the feasibility of simulating lunar regolith from available terrestrial resources.

**2.6 Validation**

To ensure the reliability of the proposed methodology, a validation step was performed by comparing the optimized Earth-based mixtures against the original lunar samples within the PCA space. The validation process involved reconstructing the lunar PCA coordinates using the computed Earth sample weights and then measuring the deviation between the reconstructed and actual lunar coordinates. This deviation was quantified using the **sum of squared residuals (SSR)**, which provided a direct measure of how closely the simulated composition matched the target.

Low SSR values indicated that the selected Earth sample mixture successfully approximated the lunar sample in the reduced PCA dimensions, while higher values suggested composition gaps or missing components in the terrestrial dataset. In addition to numerical evaluation, visual validation was performed by plotting both the lunar and reconstructed Earth-based mixtures in 2D and 3D PCA scatter plots. These plots allowed for a clear visual confirmation of whether the optimized mixtures fell within the same cluster as the target lunar samples.

This combination of quantitative and visual validation ensured that the optimization results were not only mathematically sound but also aligned with the overall objective of accurately simulating lunar regolith composition using terrestrial materials.

**3. Results and Visualisation**

**References**

Blewett, D. T., Davis, K. M., Stockstill-Cahill, K. R., & Lantz, C. (2021). Lunar simulant assessment and the limitations of X-ray diffraction in detecting amorphous phases. *Lunar and Planetary Science Conference, 52*. Lunar and Planetary Institute.

Jancsik, R., Canales, D., & Jiang, Y. (2025). Flow dynamics of lunar regolith for in-situ utilization and transportation. *Proceedings of the International Conference on Robotics and Space Engineering*. https://www.researchgate.net/publication/388075889\_Flow\_Dynamics\_of\_Lunar\_Regolith\_For\_In-Situ\_Utilization\_and\_Transportation

Kamble, V. B., & Deshmukh, S. N. (2017). Comparison between accuracy and MSE, RMSE by using proposed method with imputation technique. *Oriental Journal of Computer Science and Technology, 10*(4), 773–779. https://doi.org/10.13005/ojcst/10.04.27

Kaushik, R. (2024, October 15). KNN imputer in machine learning. *Analytics Vidhya*. https://www.analyticsvidhya.com

Morris, R. V. (1978). The surface exposure (maturity) of lunar soils: Some concepts and Is/FeO compilation. *Proceedings of the Lunar and Planetary Science Conference, 9*, 2287–2297. Lunar and Planetary Institute.

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics, 28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Tang, H., Pieters, C. M., & Keller, L. P. (2012). Simulation of nanophase iron production in lunar space-weathered materials. *Earth and Planetary Science Letters, 359–360*, 248–257. https://doi.org/10.1016/j.epsl.2012.09.039

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., … & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics, 17*(6), 520–525. https://doi.org/10.1093/bioinformatics/17.6.520

University of Central Florida. (2021). Characterization of novel lunar highland and mare simulants for ISRU research applications. *Lunar and Planetary Science Conference, 52*. Lunar and Planetary Institute.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Wang, C. (2025). Developing a high-fidelity lunar regolith simulant based on limited lunar sample availability. *Icarus*. Advance online publication. https://doi.org/10.1016/j.icarus.2025.xxx

Yan, H., Zhao, T., Li, X., & Hun, C. (2015). Synthesis of Cu-doped nano-TiO₂ by detonation method. *Ceramics International, 41*(10), 14134–14140. https://doi.org/10.1016/j.ceramint.2015.07.049

Zainal Abidin, N., Ismail, A. R., & Emran, N. A. (2018). Performance analysis of machine learning algorithms for missing value imputation. *International Journal of Advanced Computer Science and Applications, 9*(6). https://doi.org/10.14569/IJACSA.2018.090602

Zhang, H., He, Z., Lin, Y., & Hao, J. (2016). Mineralogical and chemical characteristics of Chang’E-5 lunar regolith: Insights from X-ray diffraction (XRD) and energy-dispersive X-ray spectroscopy (EDX). *Icarus, 266*, 309–318. https://doi.org/10.1016/j.icarus.2015.11.022

Zhang, J., Kring, D. A., & Zhong, J. (2015). Volcanic history of the Imbrium basin: A close-up view from impact melt-surrounded plains. *Proceedings of the National Academy of Sciences, 112*(29), 8919–8923. https://doi.org/10.1073/pnas.1503082112

(discuss about the materials Turkey/Italy)

3