# Discovering Health-Related Needs in the Community with Data Science and Open Data

Germán H. Alférez, Ph.D.
Global Software Lab, School of Engineering and Technology, Universidad de Montemorelos, Mexico
harveyalferez@um.edu.mx, www.harveyalferez.com

## Abstract

**As Seventh-day Adventists, we have been called to minister people's needs. However, discovering the real needs of people in our communities is not an easy task. At the Global Software Lab, School of Engineering and Technology, Universidad de Montemorelos, we are currently working on several research projects to analyze open data by means of data science in order to discover health-related needs in Mexico and the US. This poster summarizes four of these projects.**

## Current Projects

### 1. Application of Data Science to Classify Causes of Maternal Mortality in Mexico

- ✓ **Abstract:** In Mexico, the maternal mortality ratio is very high: 38 deaths per 100,000 live births in 2015 (WHO). The early discovery of possible causes of maternal mortality is key to reduce the number of deaths. To this end, we created a software to automatically classify patients in the following two causes of death by means of applying data science and machine learning to open data provided by the Mexican Ministry of Health: eclampsia during labor and postpartum hemorrhage [1].
- ✓ **Method:** IBM Methodology for Data Science [2].
- ✓ **Results:** The model generated with Naïve Bayes was chosen to carry out classifications within the software (precision = 0.75). The model was trained with 1,018 instances. In the case of eclampsia during labor: precision = 0.71, recall = 0.94, and $F_1$ score = 0.81. In the case of postpartum hemorrhage: precision = 0.81, recall = 0.43, and $F_1$ score = 0.56.

| Classifier | Accuracy | Precision | Recall | $F_1$ score |
|---|---|---|---|---|
| KNN | 1.00 | 0.60 | 0.59 | 0.60 |
| Logistic Regression | 0.73 | 0.73 | 0.73 | 0.73 |
| **Naïve Bayes** | **0.72** | **0.75** | **0.74** | **0.71** |
| SVM | 1.00 | 0.54 | 0.57 | 0.42 |

Table 1. Results with the generated models

### 2. Application of Data science to Confirm the Relationship between Dental Caries and Diabetes in Dental Records

- ✓ **Abstract:** In Mexico, there were over 11 million cases of diabetes in 2015 (IDF). Our contribution was to apply novel data science techniques to medical records at a dental clinic in Northeast Mexico to discover the relationship between diabetes and dental caries [3]. The analysis of data was carried out by means of unsupervised learning (K-Means). Experiments were performed on 193 dental records. Our findings corroborate the results in related work.
- ✓ **Method:** IBM Methodology for Data Science [2].
- ✓ **Results:**
  1. Diabetic patients tend to present teeth loss and food accumulation in some zones.
  2. Diabetic patients tend to present 9 to 17 teeth with caries, whereas healthy patients tend to present 1 to 9 teeth with caries.

### 3. Discovering Hidden Patterns in US Health-Related Open Data with Machine Learning

- ✓ **Abstract:** We applied unsupervised learning (K-Means) to discover hidden patterns in the Community Health Status Indicators (CHSI) open dataset provided by the Centers for Disease Control and Prevention (CDC) on Data.Gov [4]. It contains over 200 measures for each of the 3,141 US counties.
- ✓ **Method:** IBM Methodology for Data Science [2].
- ✓ **Results:**
  1. An increasing number of primary care physicians/dentists per county correlates with a decreasing number of people with diabetes per county.
  2. Increasing numbers of Medicare beneficiaries and community/migrant health centers per county correlate with a decreasing number of people with diabetes per county.

### 4. Discovering Mission-Oriented Patterns with Open Data in New York City

- ✓ **Abstract:** We analyzed an open dataset of motor vehicle collisions in NYC, which is freely provided by NYPD. This dataset registers vehicle collisions in Bronx, Brooklyn, Manhattan, Queens, and Staten Island from 2014 to 2016. It contains 932,904 registered incidents. Each registered incident has 30 variables. We applied the K-Means algorithm to this dataset [5]. We have a big interest on applying data science to understand the needs of people in NYC and take action. See our previous results on Adventist Review [6].
- ✓ **Method:** IBM Methodology for Data Science [2].
- ✓ **Results:**
  1. On Thursdays, Fridays, and Saturdays, drivers tend to drive aggressively. This situation increases the number of accidents during those days.
  2. On Fridays, around Prospect Park, Brooklyn, there were over 77,000 pedicab accidents.

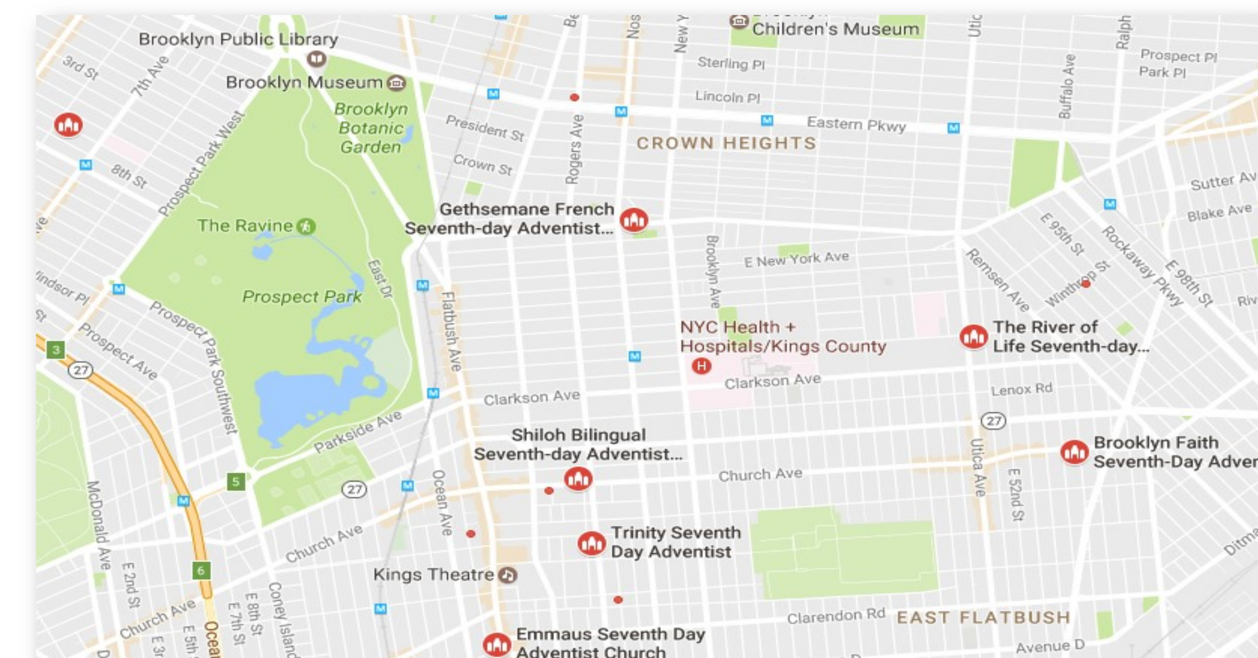


Figure 1. SDA churches near Prospect Park

## Future Work

We are planning to apply data science to data from Adventist World Radio and the North American Division. We are currently finishing a project to understand the needs of people in the 10/40 Window.

1. R. Domínguez, "Aplicación de ciencia de datos para la creación de software predictivo de morbimortalidad materna en México," M.S. thesis, School of Eng. and Tech., Universidad de Montemorelos, Mexico, 2017.
2. *Foundational Methodology for Data Science*, IBM, Somers, NY, 2015.
3. G.H. Alférez et al. "Application of data science to discover the relationship between dental caries and diabetes in dental records," presented at the 2016 International Conference on Health Informatics and Medical Systems (HIMS 2016), Las Vegas, NV, 2016.
4. G.H. Alférez. (2016, November 11). *Discovering hidden patterns in US health-related open data with machine learning* [Online]. Available: http://www.sdadata.org/blog/discovering-hidden-patterns-in-us-health-related-open-data-with-machine-learning
5. G.H. Alférez. (2016, May 1). *Discovering Mission-Oriented Patterns with Open Data in New York City* [Online]. Available: http://www.sdadata.org/blog/discovering-mission-oriented-patterns-with-open-data-in-new-york-city
6. G.H. Alférez, "Tweeting in New York City - data science can teach us to sympathize," Adventist Review, vol. 193, no. 2, pp.47-49, Feb. 2016.

# AUTOMATIC CLASSIFICATION OF PLUTONIC ROCKS WITH MACHINE LEARNING APPLIED TO EXTRACTED COLORS ON iOS DEVICES

Sarah Hernández[1]; Germán H. Alférez, Ph.D.[1]; Benjamin L. Clausen, Ph.D.[2]; Ana M. Martínez, Ph.D.[2]
[1] School of Engineering and Technology of Montemorelos University; [2] Department of Earth and Biological Sciences of Loma Linda University

# INTRODUCTION

Plutonic rocks are formed when magma cools and solidifies below the Earth's surface [1]. Lightness and color are properties used for the classification of plutonic rocks; however, these attributes can be difficult to describe because perceived rock colors depend on the observer's experience [2]. Moreover, although the classification of plutonic rocks can be done using data from various instrumental techniques, these approaches tend to be expensive and time-consuming. This research extracts dominant shades and colors from plutonic rock images to train several machine learning algorithms and deploy the best model in an iOS app for the automatic classification of four classes of plutonic rocks in order from darker to lighter: gabbro, diorite, granodiorite, and granite.

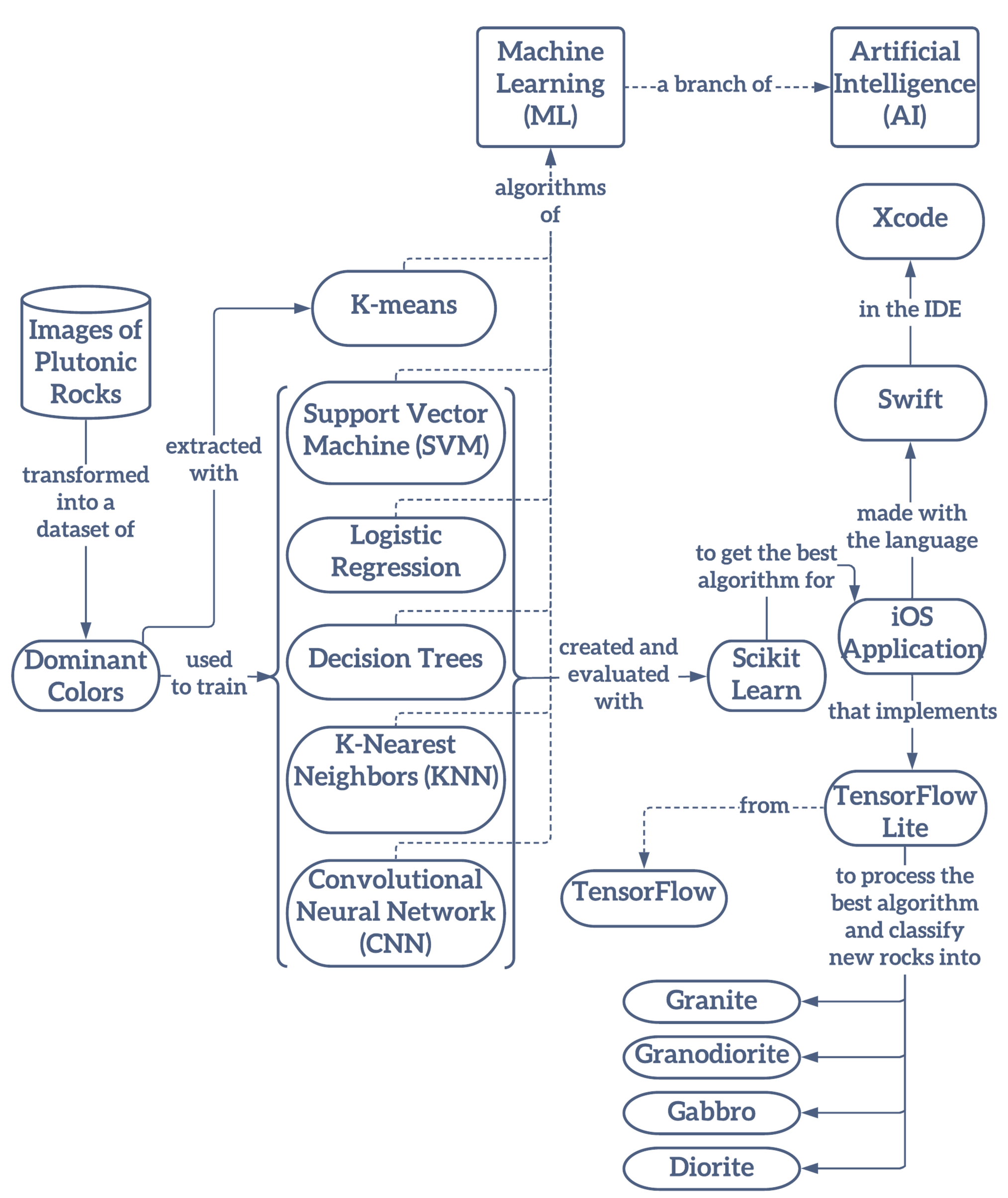

Figure 1. Underpinnings of our approach

| Author | Year | Dataset | Algorithm of extraction | Extracted features | Dataset for training | Best trained model | Results | Results in the field |
|---|---|---|---|---|---|---|---|---|
| E. Vázquez and H. Alférez [3] | 2021 | 7846 patches from 71 images of six plutonic rock types (diorite, gabbro, granite, granodiorite, monzodiorite, and tonalite) | | | 50% of the patches | 4-layer structure CNN | Accuracy: 95%, Precision: 96%, Recall: 95%, F1-score: 95% | Android application accuracy results: 70% for gabbro, 28.5% for diorite, 100% for granodiorite, and 28.5% for granite |
| Fan et al. [4] | 2020 | 3,208 images of 28 rock lithology categories | | | 80% of the images | SqueezeNet | Model accuracy: 94.55% | Android application accuracy: 94.55% |
| Fan et al. [5] | 2020 | 3,795 images of 30 different rock types | | | 80% of the images | ShuffleNet | Model accuracy: 95.30% | Android application accuracy: 97.65% |
| Ran et al. [6] | 2019 | 14,589 patches from 2,290 images of six rock types (granite, limestone, conglomerate, sandstone, shale, mylonite) | | | 60% of the patches | 3-layer structure CNN | Model accuracy: 97.76% | |
| Zhang et al. [7] | 2019 | 481 images of four minerals (K-feldspar, perthite, plagioclase, and quartz) | Inception-v3 CNN model | High-level features (such as chromatic aberration and texture) | 90% of extracted features | Stacking model (LR, SVM, and MLP) | Model accuracy: 90.9% | |
| Maitre et al. [8] | 2019 | 3,192 sub-images of the view taken to a surface with 27 mineral grains species | Simple Linear Iterative Clustering (SLIC) algorithm | Color and Peak intensity of super-pixel histograms | 70% of extracted features | Random Forest (RF) | Model accuracy: 89% | |
| Cheng and Guo [9] | 2017 | 4,200 images of three feldspar sandstones rock types (coarse, medium granular, and fine) | | | 75% of the images | 6-layer structure CNN | Model accuracy: 98.5% | |

Table 1. Related work

# METHODOLOGY

We used pictures from plutonic rocks that were classified by using petrography and chemistry data to train the models [3].

## 1. Color extraction

The dominant colors of plutonic rock images were extracted with the k-means algorithm by grouping the image pixels according to the RGB and CIELAB color spaces (see Fig. 2).

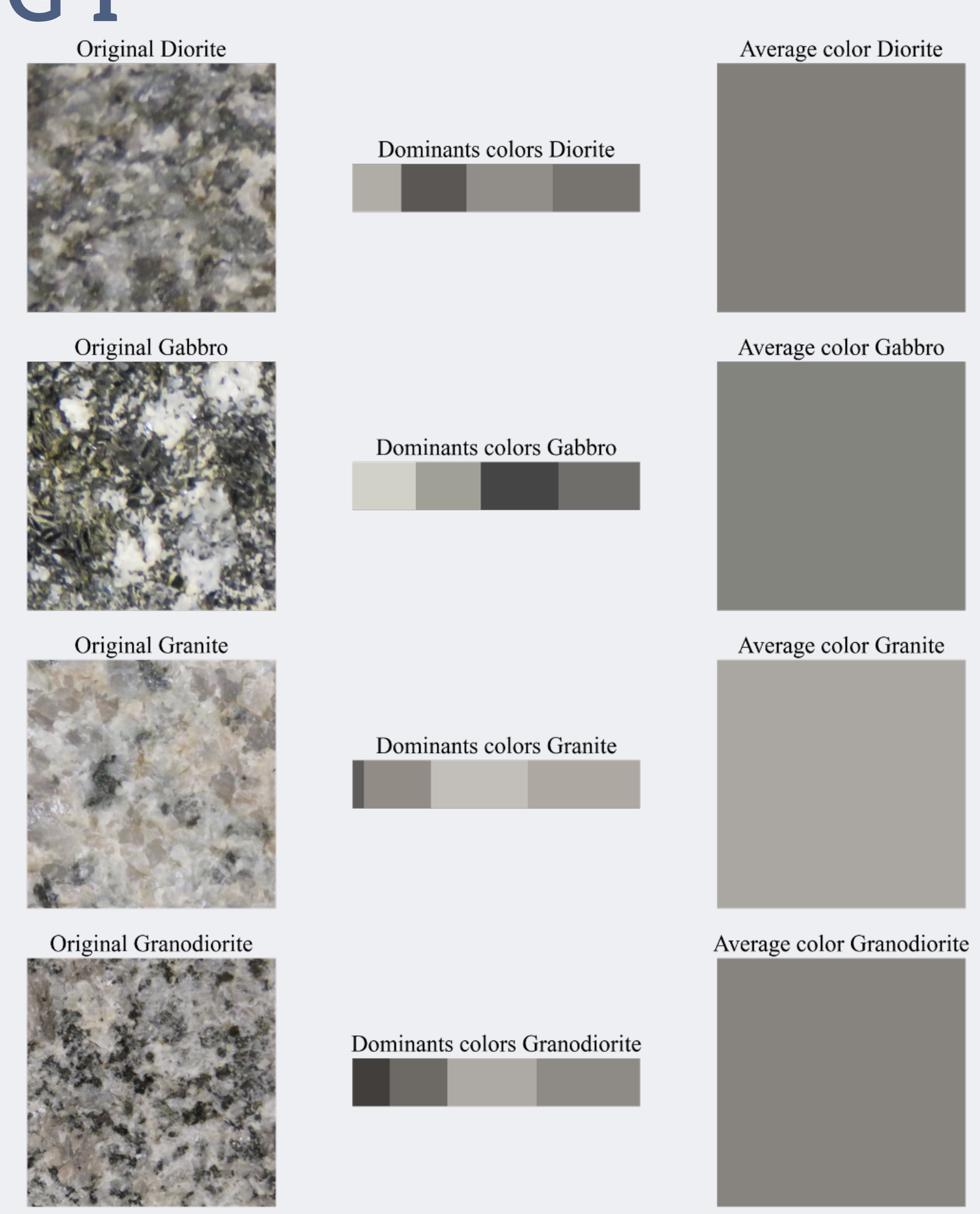

Figure 2. Dominant colors in sample images

## 2. Model training and evaluation

The data of the four dominant colors were used to create and evaluate several machine learning models with the following algorithms: Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Trees (DT), Support Vector Machine (SVM), and Convolutional Neural Networks (CNN). The best results during validation were for the model generated with the K-Nearest Neighbors trained with 283 images in the CIELAB color space (see Fig. 3). Results gave accuracy, precision, recall, and F-score average values of 93%.

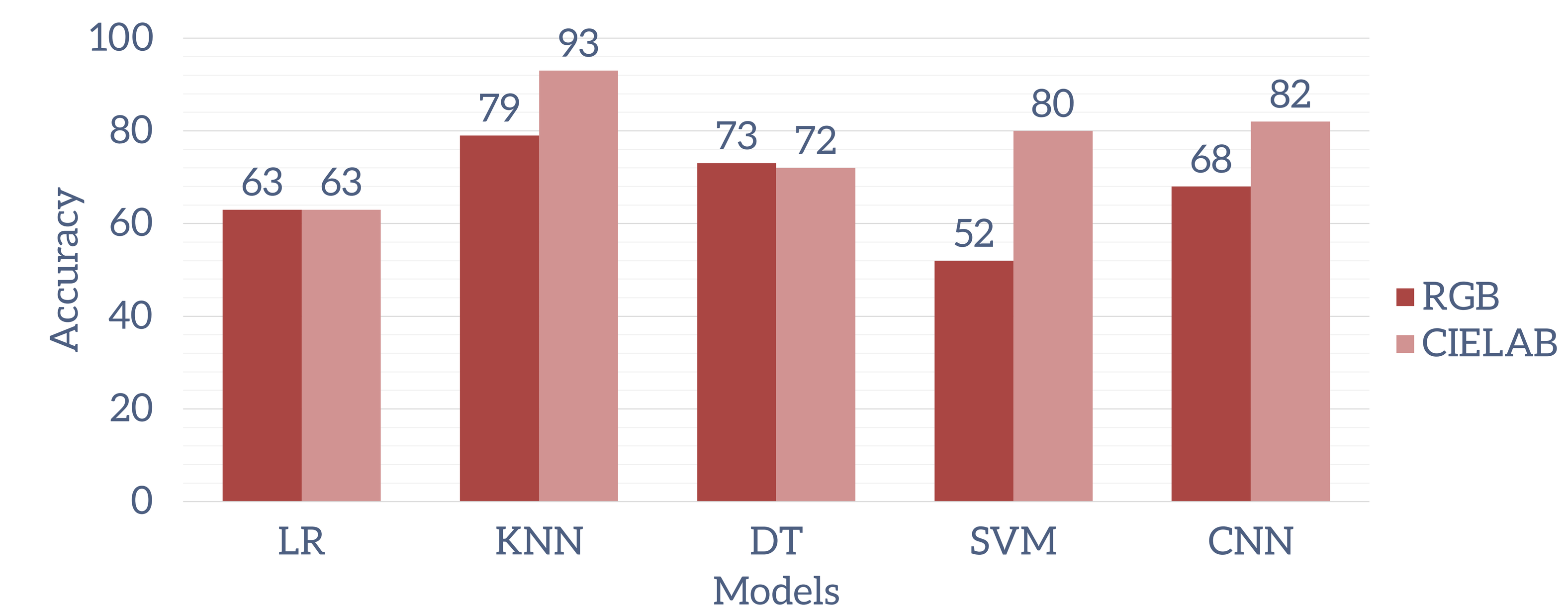

Figure 3. Accuracy of the models in RGB and CIELAB formats

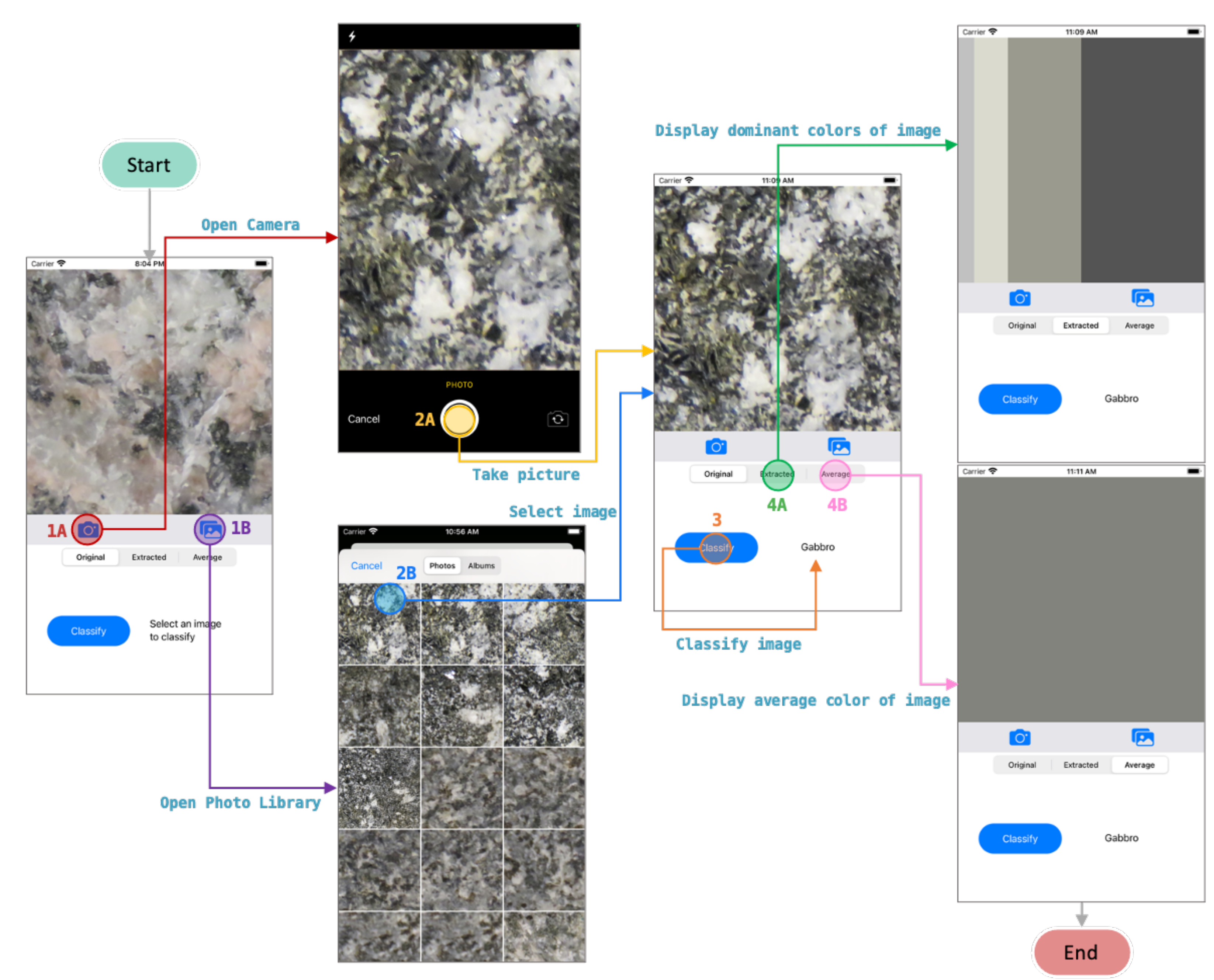

Figure 4. Application workflow

# RESULTS

The KNN model was deployed after validation on an iOS application that classifies the extracted colors in new images of the four rock types (see Fig. 4). The application was tested in the field with 34 images, and the following average accuracy results were obtained: 70% for gabbro, 28.5% for diorite, 20% for granodiorite, and 85.7% for granite.

| Class | Taken images | Correctly classified | |
|---|---|---|---|
| | | Images | Percentage |
| Granite | 7 | 6 | 85.7% |
| Granodiorite | 10 | 2 | 20% |
| Gabbro | 10 | 7 | 70% |
| Diorite | 7 | 2 | 28.5% |
| Total | 34 | 17 | 50% |

Table 2. Accuracy results of application evaluation

# CONCLUSIONS

The high accuracy when classifying gabbro samples was because they are noticeable darker than samples of the other 3 classes. Similarly, granites were noticeably lighter. In contrast, diorite and granodiorite share characteristics of the other rock types closest to them in the dark-light sequence; therefore, it is more difficult to automatically classify them based on their dominant colors.

# REFERENCES

[1] f. y. n., RACEFEN Glosario de geología, Spanish, Access date: 01/04/2021, España: Real academia de ciencias exactas, físicas y naturales.
[2] Natural Resources Conservation Service, "Part 631: Geology," in National Engineering Handbook, 210-VI, Access date: 12/16/2020, 2012, ch. 4, p. 7.
[3] E. Vázquez and H. Alférez, "Using Deep Learning for Automatic Classification of Plutonic Rocks with Mobile Devices", 2021.
[4] G. Fan, F. Chen, D. Chen, and Y. Dong, "Recognizing Multiple Types of Rocks Quickly and Accurately Based on Lightweight CNNs Model," *IEEE Access*, vol. 8, pp. 55 269–55 278, 2020. DOI: 10.1109/access.2020. 2982017.
[5] G. Fan, F. Chen, D. Chen, Y. Li, and Y. Dong, "A Deep Learning Model for Quick and Accurate Rock Recognition with Smartphones," *Mobile Information Systems*, vol. 2020, pp. 1–14, 2020. DOI: 10.1155/2020/7462524.
[6] X. Ran, L. Xue, Y. Zhang, Z. Liu, X. Sang, and J. He, "Rock Classification from Field Image Patches Ana- lyzed Using a Deep Convolutional Neural Network," *Mathematics*, vol. 7, no. 8, p. 755, 2019. DOI: 10.3390/math7080755.
[7] Y. Zhang, M. Li, S. Han, Q. Ren, and J. Shi, "Intelligent Identification for Rock-Mineral Microscopic Images Using Ensemble Machine Learning Algorithms," *Sensors*, vol. 19, no. 18, p. 3914, 2019. DOI: 10 .3390/s19183914.
[8] J. Maitre, K. Bouchard, and L. P. Bédard, "Mineral grains recognition using computer vision and machine learning," *Computers & Geosciences*, vol. 130, pp. 84– 93, 2019. DOI: 10.1016/j.cageo.2019.05.009.
[9] G. Cheng and W. Guo, "Rock images classification by using deep convolution neural network," *Journal of Physics: Conference Series*, vol. 887, p. 012 089, 2017. DOI: 10.1088/1742-6596/887/1/012089.

# AUTOMATIC CLASSIFICATION OF PLUTONIC ROCKS WITH MACHINE LEARNING APPLIED TO EXTRACTED COLORS ON iOS DEVICES

Sarah Hernández[1]; Germán H. Alférez, Ph.D.[1]; Benjamin L. Clausen, Ph.D.[2]; Ana M. Martínez, Ph.D.[2]

[1] School of Engineering and Technology of Montemorelos University; [2] Department of Earth and Biological Sciences of Loma Linda University

## INTRODUCTION

Plutonic rocks are formed when magma cools and solidifies below the Earth's surface [1]. Lightness and color are properties used for the classification of plutonic rocks; however, these attributes can be difficult to describe because perceived rock colors depend on the observer's experience [2]. Moreover, although the classification of plutonic rocks can be done using data from various instrumental techniques, these approaches tend to be expensive and time-consuming. This research extracts dominant shades and colors from plutonic rock images to train several machine learning algorithms and deploy the best model in an iOS app for the automatic classification of four classes of plutonic rocks in order from darker to lighter: gabbro, diorite, granodiorite, and granite.

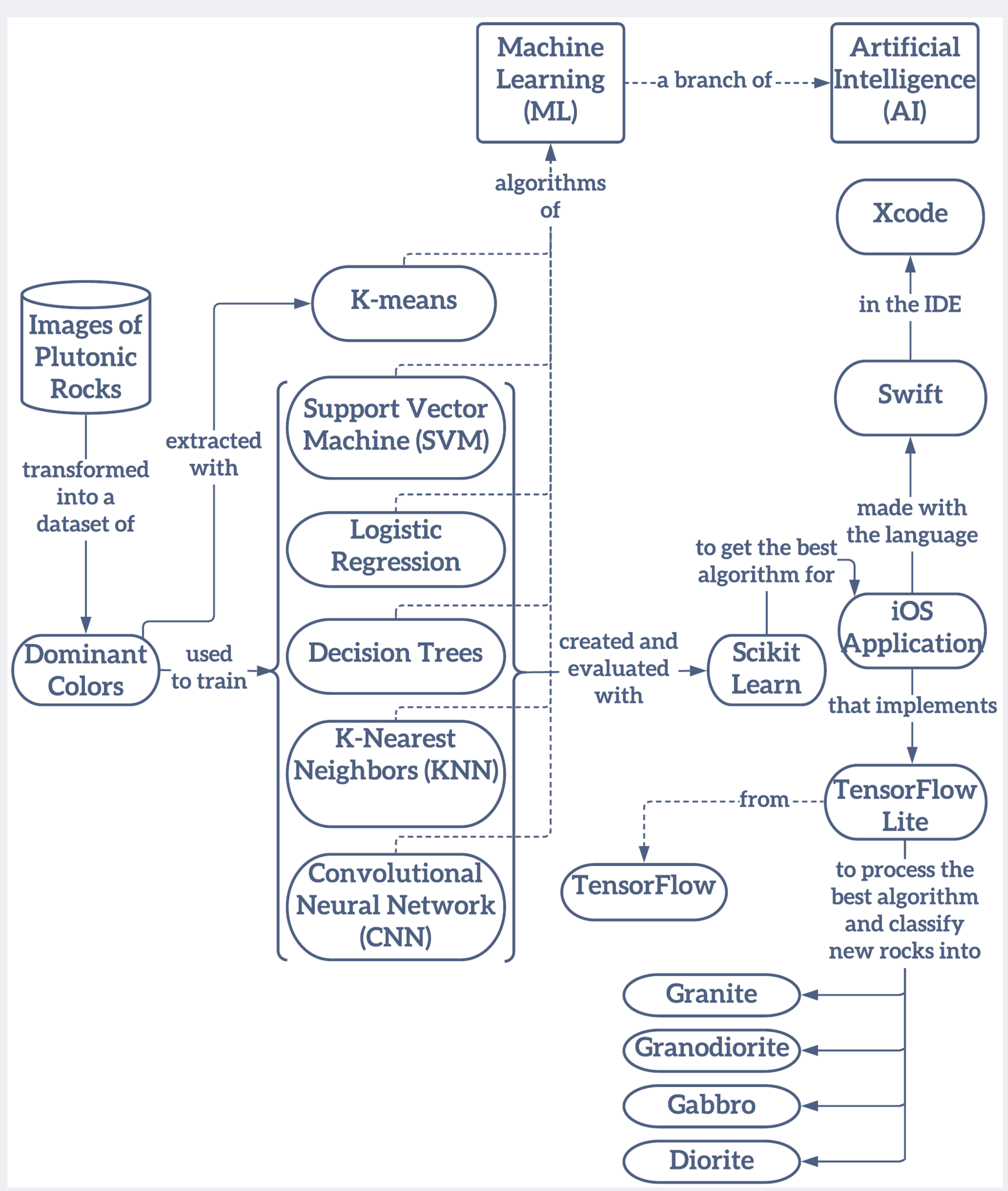


Figure 1. Underpinnings of our approach

| Author | Year | Dataset | Algorithm of extraction | Extracted features | Dataset for training | Best trained model | Results | Results in the field |
|---|---|---|---|---|---|---|---|---|
| E. Vázquez and H. Alférez [3] | 2021 | 7846 patches from 71 images of six plutonic rock types (diorite, gabbro, granite, granodiorite, monzodiorite, and tonalite) | | | 50% of the patches | 4-layer structure CNN | Accuracy: 95%, Precision: 96%, Recall: 95%, F1-score: 95% | Android application accuracy results: 70% for gabbro, 28.5% for diorite, 100% for granodiorite, and 28.5% for granite |
| Fan et al. [4] | 2020 | 3,208 images of 28 rock lithology categories | | | 80% of the images | SqueezeNet | Model accuracy: 94.55% | Android application accuracy: 94.55% |
| Fan et al. [5] | 2020 | 3,795 images of 30 different rock types | | | 80% of the images | ShuffleNet | Model accuracy: 95.30% | Android application accuracy: 97.65% |
| Ran et al. [6] | 2019 | 14,589 patches from 2,290 images of six rock types (granite, limestone, conglomerate, sandstone, shale, mylonite) | | | 60% of the patches | 3-layer structure CNN | Model accuracy: 97.76% | |
| Zhang et al. [7] | 2019 | 481 images of four minerals (K-feldspar, perthite, plagioclase, and quartz) | Inception-v3 CNN model | High-level features (such as chromatic aberration and texture) | 90% of extracted features | Stacking model (LR, SVM, and MLP) | Model accuracy: 90.9% | |
| Maitre et al. [8] | 2019 | 3,192 sub-images of the view taken to a surface with 27 mineral grains species | Simple Linear Iterative Clustering (SLIC) algorithm | Color and Peak intensity of super-pixel histograms | 70% of extracted features | Random Forest (RF) | Model accuracy: 89% | |
| Cheng and Guo [9] | 2017 | 4,200 images of three feldspar sandstones rock types (coarse, medium granular, and fine) | | | 75% of the images | 6-layer structure CNN | Model accuracy: 98.5% | |

Table 1. Related work

## METHODOLOGY

We used pictures from plutonic rocks that were classified by using petrography and chemistry data to train the models [3].

### 1. Color extraction

The dominant colors of plutonic rock images were extracted with the k-means algorithm by grouping the image pixels according to the RGB and CIELAB color spaces (see Fig. 2).

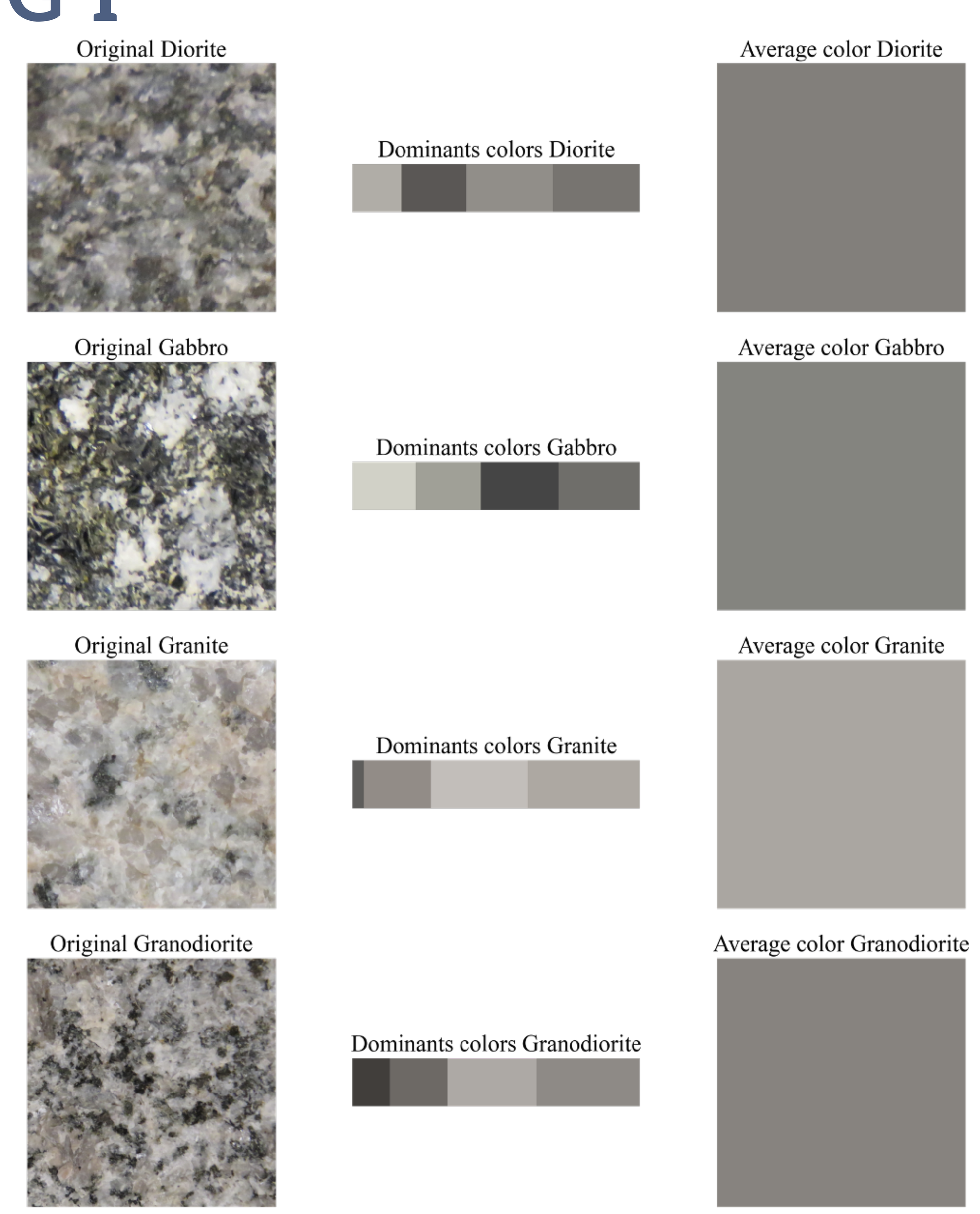


Figure 2. Dominant colors in sample images

### 2. Model training and evaluation

The data of the four dominant colors were used to create and evaluate several machine learning models with the following algorithms: Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Trees (DT), Support Vector Machine (SVM), and Convolutional Neural Networks (CNN). The best results during validation were for the model generated with the K-Nearest Neighbors trained with 283 images in the CIELAB color space (see Fig. 3). Results gave accuracy, precision, recall, and F-score average values of 93%.

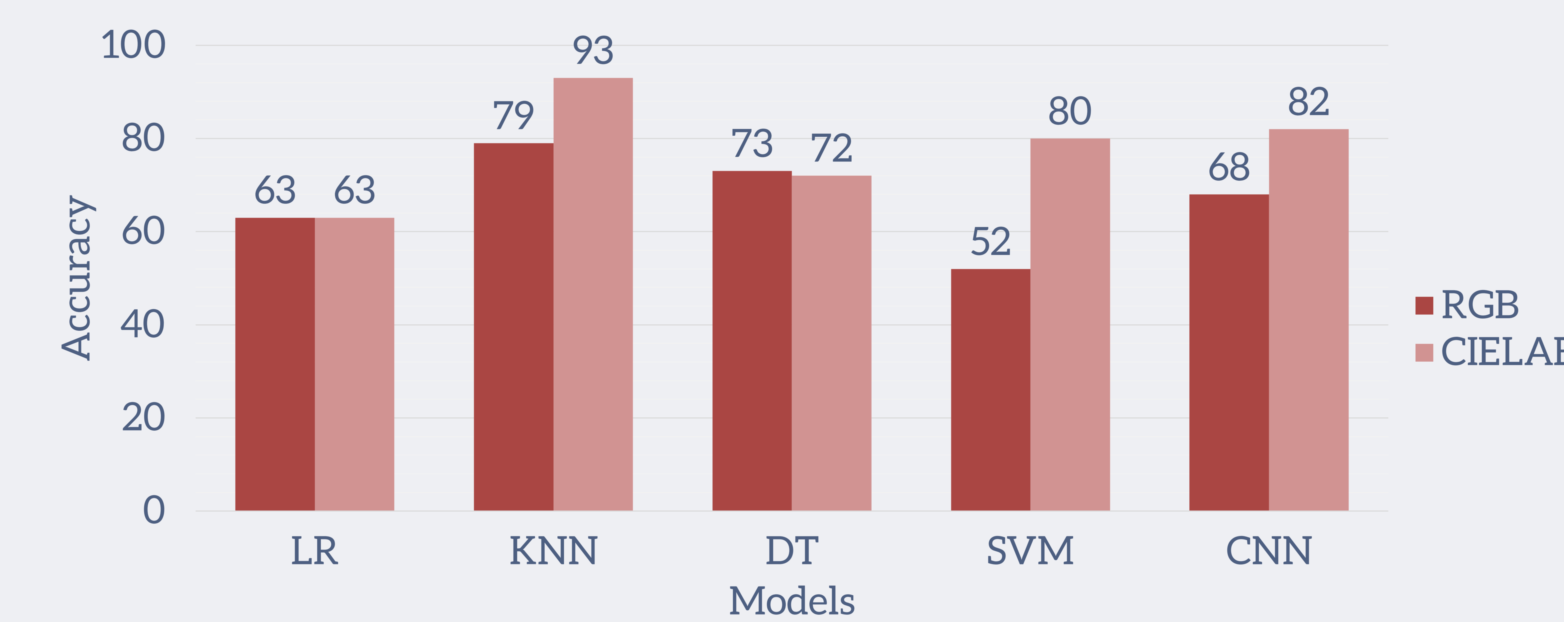


Figure 3. Accuracy of the models in RGB and CIELAB formats

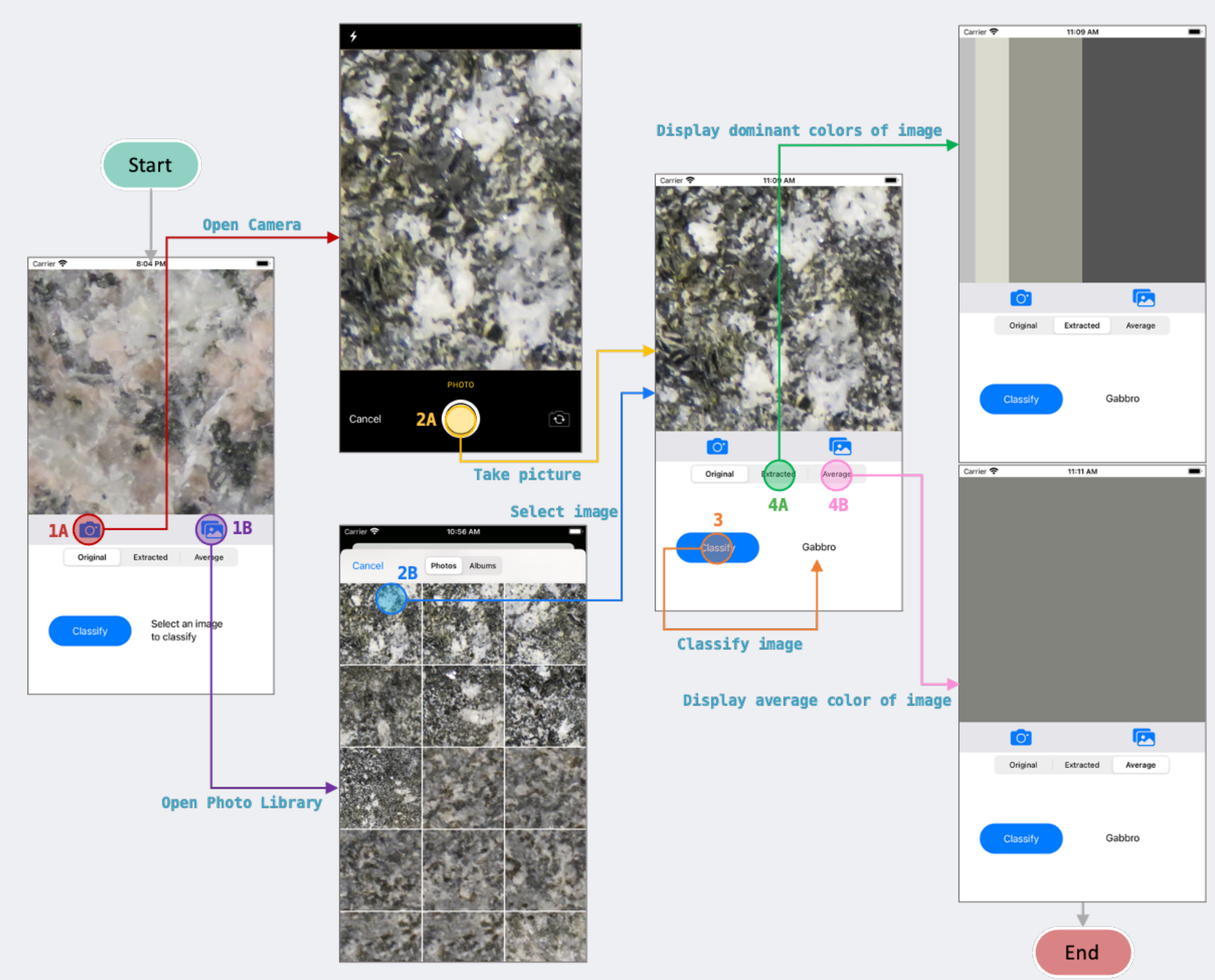


Figure 4. Application workflow

## RESULTS

The KNN model was deployed after validation on an iOS application that classifies the extracted colors in new images of the four rock types (see Fig. 4). The application was tested in the field with 34 images, and the following average accuracy results were obtained: 70% for gabbro, 28.5% for diorite, 20% for granodiorite, and 85.7% for granite.

| Class | Taken images | Correctly classified | |
|---|---|---|---|
| | | Images | Percentage |
| **Granite** | 7 | 6 | 85.7% |
| **Granodiorite** | 10 | 2 | 20% |
| **Gabbro** | 10 | 7 | 70% |
| **Diorite** | 7 | 2 | 28.5% |
| **Total** | **34** | **17** | **50%** |

Table 2. Accuracy results of application evaluation

## CONCLUSIONS

The high accuracy when classifying gabbro samples was because they are noticeable darker than samples of the other 3 classes. Similarly, granites were noticeably lighter. In contrast, diorite and granodiorite share characteristics of the other rock types closest to them in the dark-light sequence; therefore, it is more difficult to automatically classify them based on their dominant colors.

## REFERENCES

[1] f. y. n., RACEFEN Glosario de geología, Spanish, Access date: 01/04/2021, España: Real academia de ciencias exactas, físicas y naturales.
[2] Natural Resources Conservation Service, "Part 631: Geology," in National Engineering Handbook, 210-VI, Access date: 12/16/2020, 2012, ch. 4, p. 7.
[3] E. Vázquez and H. Alférez, "Using Deep Learning for Automatic Classification of Plutonic Rocks with Mobile Devices", 2021.
[4] G. Fan, F. Chen, D. Chen, and Y. Dong, "Recognizing Multiple Types of Rocks Quickly and Accurately Based on Lightweight CNNs Model," *IEEE Access*, vol. 8, pp. 55 269–55 278, 2020. DOI: 10.1109/access.2020. 2982017.
[5] G. Fan, F. Chen, D. Chen, Y. Li, and Y. Dong, "A Deep Learning Model for Quick and Accurate Rock Recognition with Smartphones," *Mobile Information Systems*, vol. 2020, pp. 1–14, 2020. DOI: 10.1155/2020/7462524.
[6] X. Ran, L. Xue, Y. Zhang, Z. Liu, X. Sang, and J. He, "Rock Classification from Field Image Patches Ana- lyzed Using a Deep Convolutional Neural Network," *Mathematics*, vol. 7, no. 8, p. 755, 2019. DOI: 10.3390/math7080755.
[7] Y. Zhang, M. Li, S. Han, Q. Ren, and J. Shi, "Intelligent Identification for Rock-Mineral Microscopic Images Using Ensemble Machine Learning Algorithms," *Sensors*, vol. 19, no. 18, p. 3914, 2019. DOI: 10 .3390/s19183914.
[8] J. Maitre, K. Bouchard, and L. P. Bédard, "Mineral grains recognition using computer vision and machine learning," *Computers & Geosciences*, vol. 130, pp. 84– 93, 2019. DOI: 10.1016/j.cageo.2019.05.009.
[9] G. Cheng and W. Guo, "Rock images classification by using deep convolution neural network," *Journal of Physics: Conference Series*, vol. 887, p. 012 089, 2017. DOI: 10.1088/1742-6596/887/1/012089.