Journal of Building Engineering 29 (2020) 101175

Contents lists available at ScienceDirect

Journal of Building Engineering

journal homepage: http://www.elsevier.com/locate/jobe

Predicting the suitability of lateritic soil type for low cost sustainable housing with image recognition and machine learning techniques

Tuza A. Olukan [a,b], Yu-Cheng Chiou [a], Cheng Hsiang Chiu [a], Chia-Yun Lai [a], Sergio Santos [a], Matteo Chiesa [a,b,*]

[a] Department of Physics and Technology, UiT The Arctic University of Norway, 9010, Tromsø, Norway
[b] Laboratory for Energy and NanoScience (LENS), Khalifa University of Science and Technology, Masdar Campus, PO Box 54224, Abu Dhabi, United Arab Emirates

ARTICLE INFO

Keywords:
Increase
Sustainability
Building materials
Housing
Shortage

ABSTRACT

From a sustainability point of view, laterites-compressed earth bricks (LCEB) are a promising substitute for building structures in place of the conventional concrete masonry units. On the other hand, techniques for identifying and classifying laterites soil for compressed earth bricks (CEB) production are still relying on direct human expertise or 'experts'. Human experts exploit direct visual inspection and other basic senses such as smelling, touching or nibbling to generate a form of binomial classification, i.e. suitable or unsuitable. The source of predictive power is otherwise supposed to be found in color, scent, texture or combinations of these. Lack of clarity regarding the actual method and the possible explanatory mechanisms lead to 1) difficulties to train other people into the skills and 2) might also add to apathy to using CEB masonry units for housing. Here we systematize the selection method of experts. We chose imaging analysis techniques based on 1) easiness in image acquisition (Digital Camera) and 2) availability of machine learning and statistical techniques. We find that most of the predictive power of the 'expert' can be packed into visual inspection by demonstrating that with image analysis alone we get a 98% match. This makes it practically unnecessary the study of any other 'expert' skills and provides a method to alleviate the housing problems dealing with material construction in the developing world.

1. Introduction

Meeting the housing need of the expanding human population, especially in the developing economies, with the current building technology (concrete masonry unit) is not only prohibitively expensive but environmentally unsustainable. Here, these two parameters define our investigation. In short we are seeking 1) cheap and abundant materials that are also available in the regions of interest and 2) materials that do not involve complex or overwhelming post-processing such as extra refining or the needs of additives. In this respect, laterites-compressed earth bricks (LCEB) technology offers a cheaper, simple, and more sustainable alternative to current technology [1]. CEB units are simple masonry elements obtained by compaction of humid soil types (usually laterites) with the option of a chemical binder [2]. The presence of clay in the soil is advantageous since clay adheres the soil grains together hence eliminating or limiting the usage of a chemical stabilizer. Since laterites naturally contain clay materials and sesquioxides, these are our preferred choice materials for CEB units production [3]. Laterites further meet our second requirement since the material is locally sourced and widely available in the subsoil of inter-tropical regions of all the world continents [4].

The main disadvantage of laterites for LCEB relates to variability, namely, the mineralogical composition of the soil varies by geographical location [4,5] making it challenging to generalize their engineering behavior. In principle however, this is a technical problem, and more thoroughly, a problem of soil selection that should not highly affect cost and that has no effect on sustainability or transportation. Thus, we propose to reduce our problem to a classification problem, that is, to the selection of an appropriate machine learning technique for classification purposes.

2. Background and motivation

Engineering methodologies to deal with soil properties, variance and

* Corresponding author. Department of Physics and Technology, UiT The Arctic University of Norway, 9010, Tromsø, Norway.
E-mail address: matteo.chiesa@ku.ac.ae (M. Chiesa).

https://doi.org/10.1016/j.jobe.2020.101175
Received 17 October 2019; Received in revised form 22 November 2019; Accepted 5 January 2020
Available online 10 January 2020

overall variation are already available. For example, particle size distributions (PSD) and plasticity tests sufficiently predict the engineering properties of most soils. Nevertheless engineering indices fail to predict the field behavior of laterites accurately [6]. It has been reported that the properties of laterites are influenced significantly by the pretest and sample preparation procedures. This leads to inconsistent results since what is theoretically expected and predicted is not met during practical implementations [7–11]. Some authors tried addressing these inconsistencies by classifying laterites based on the parent material, the degree of weathering or by exploiting commonly employed engineering classification systems. Such efforts however, have found little acceptance in the soil engineering community presumably because of lack of reproducibility due to perturbations to properties during testing, amongst other [9,12–15]. In addition to the core articles referenced, there are several studies in the literature (both published and unpublished) targeted at addressing the localized problem of laterites classification for engineering use in restricted areas. Results from such studies however, are usually incomplete and tend to have conflicting viewpoints [12,16]. Despite the inconsistency reported in the literature on the selection of lateritic soil for engineering use, normative documents on this subject exist with recommended techniques. For instance, according to the **ARS 680:1996** (Compressed earth blocks - code of practice for the production of compressed earth blocks) [17], it is recommended that soil selection techniques should either be based on the user's empirical knowledge or a set of laboratory tests procedures.

Ultimately, taking word of mouth and our personal experience with laterites for engineering construction as a reference - and otherwise considering the lack of literature and reported documentation in this respect – the actual selection of the material in most developing country is mainly based on empirical knowledge or the "intuitive skills" of human experts. This could partly due to the facts that soil testing laboratories are few or far between in this region of the world, making it easier, more convenient and cheaper to rely on the experiences of the "experts". From now on we refer to these human experts as "experts". Experts use techniques such as visual inspection, touching, smelling and nibbling as described in Ref. [18] to judge the suitability of laterite for CEB making. Out of these methods, it is unclear which contains the predictive power (color, scent, or texture). We believe such situation of ambiguity of method, together with the uncertainty in terms of the feature, or set of features, carrying most predictive power, to be a main reason behind the difficulties to train other people into the skillset of current experts. The problem of unreliability and lack of clear scientific standards might further explain the apathy toward CEB masonry units for housing.

To address these challenges, we propose a simple, computationally fast, and efficient method. By using a combination of image techniques, machine learning, and statistics for classification of laterites for CEB suitability using soil color as the sole feature, we propose to capture "most" of the predictive power of the expert. Our choice for soil color as the main predictive feature is predicated on the fact that 1) it has been an attribute previously and extensively used in determining soil characteristics in the literature [19–27], and 2) on the current advance and reduced cost of machine learning techniques exploiting image processing for general classification [28,29]. For instance, Liles et al. developed a predictive soil model using quantitative color measurements to established a trend between soil color and soil organic matter [30]. Similar work was done by Viscarra Rosel et al. in Refs. [31–33] using a digital camera. To avoid the lighting effects and calibration, some authors monitor their samples under controlled illumination and color chips [34–38] while there are also studies in the literature detailing the use of image analysis techniques and machine learning in classifying soil based on their textural composition [26,39–47], the methodology deployed in those works is usually overcomplicated and computationally intensive for our purpose.

We next set to structure the problem of laterites identification for CEB production as a binary classification problem, i.e. for a given laterite image, we seek to develop a classifier that assigns each image a 'Suitable' or 'Unsuitable' label. A supervised, i.e. labeled data for training purposes, algorithm is viable since we have available images labeled by experts and it is precisely the skills of the expert that we want to duplicate. We note that we use the term "training set" here relatively loosely but precisely. In short, with training data and with training we mean the data that is used to generate a generic model based on statistical parameters to classify data and with training the fitting of the values accordingly. The problem relating to the actual predictive power of the expert is not considered here and we assume that the expert is 100% reliable while acknowledging that this is defined in this way for convenience and overestimates human skills. The classification algorithms and techniques deployed in this work are summarized in Fig. 1.

3. Sampling method

First the images were acquired in an in-house setup system for all the datasets used in this work. The images were pre-processed by converting them from RGB to HSV color space for efficient computation. A color histogram is afterwards used to extract the color features from the images and stored as a vector in a dataset. We then use cosine similarity measurements and a statistical technique, fully described elsewhere [48], that applies even when the central limit theorem does not apply. Our technique is relatively simple and has the advantage to force reproducibility by considering a much larger set of sampling data than it is typically necessary in the use of normally distributed data to work out average values. Specifically, this statistical method allows us to determine the minimum sample size required to represent a given population without compromising accuracy and by forcing reproducibility at the expense of requiring a larger sample set. Data availability for training is not a problem in our case since we find that only a small fraction of the images (about 5%) are required as training datasets to successfully generate a valid classification algorithm. Furthermore, the method comes with the advantage of generating a metric that acts as a threshold (decision boundary line) to determine contrast heterogeneity, provided it exists, between the query image and the predictive algorithm generated via the training set. In summary, we use this method as a classifier to generate the training sets used to assigning the appropriate labels to the images (feature vectors).

4. Materials and methods

4.1. Soil sample collection

We sourced ten different laterite samples from different locations in Nigeria. With the help of the experts, only three samples were deemed suitable for CEB production. For further analysis, we selected three samples from both classes ("Suitable" and "Unsuitable"). For robustness and in order to account for cost, we quantify the suitability of any laterite by the relative material cost used in producing CEB masonry units (C_{CEB}) from the laterite compared to the material cost in producing a concrete masonry unit (C_{CMU}) of equivalent strength. All things been equal, we consider a laterite sample as suitable when $C_{CEB} < C_{CMU}$ (see Fig. 2). This clarification is necessary because technically, even the laterites considered "Unsuitable' can be enhanced for CEB masonry elements with the right treatment. We note however that the treatment cost of enhancing such laterite, unsuitable laterite, for CEB uses will often run against our seeking for the advantages of selecting it in the first place over the conventional concrete units.

Even though, we are more concerned with systemizing the predictive power of the experts, for completeness we performed some laboratory techniques in order to 1) obtain a first estimation of the variation of all the laterites samples, 2) test available methods independently in this work and 3) compare such results with the results obtained via our approach. Specifically, we performed the sedimentation test as detailed in Refs. [18,49] to find the textural composition of all the laterite

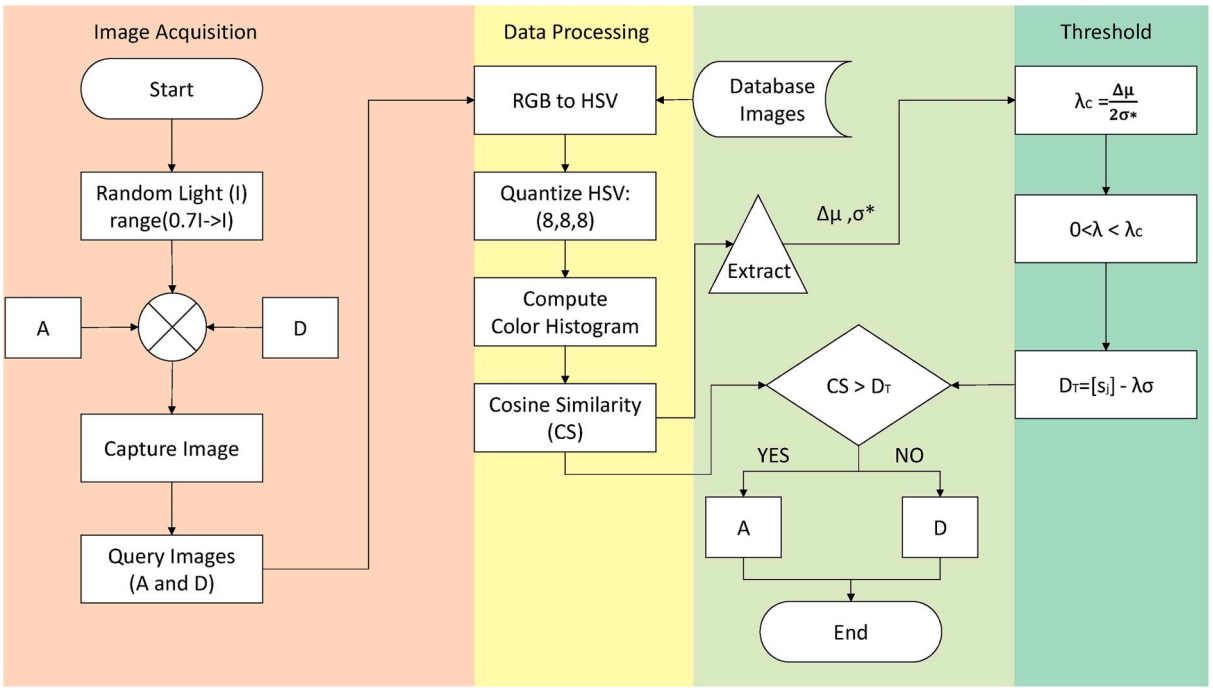

Fig. 1. Flow chart process illustrating the soil classification procedures.

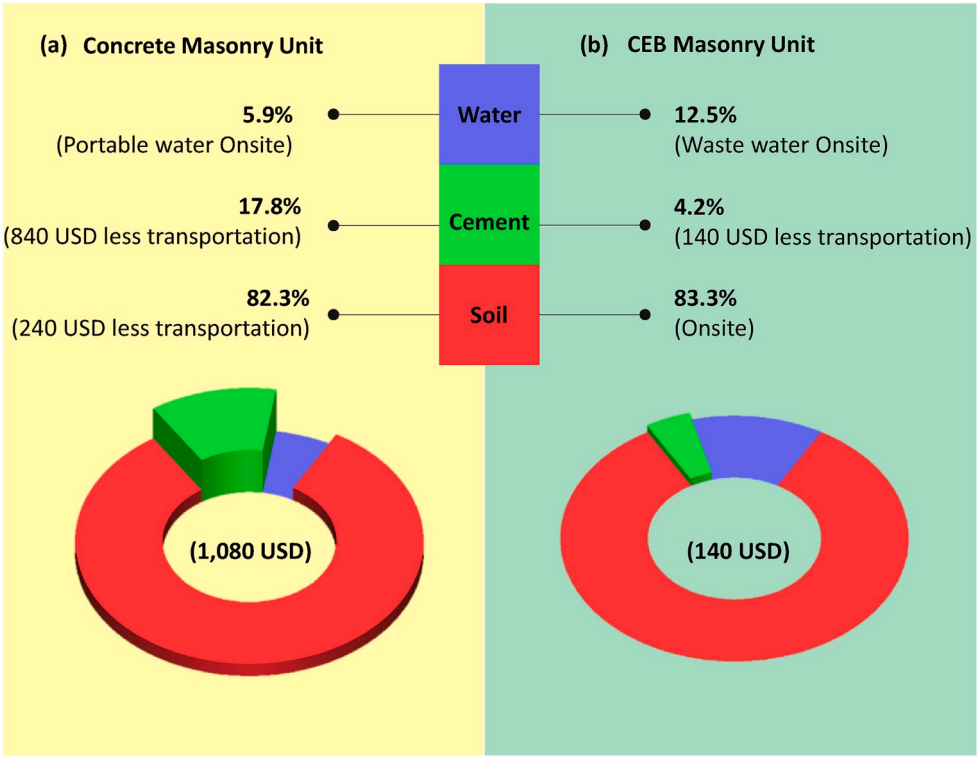

Fig. 2. The estimated material cost of masonry units required to build a single-story unit in Nigeria.

samples (see Table 1).

4.2. Image acquisition

The images were all acquired in an in-house fabricated setup (Fig. 3). The setup consisted of a DSLR Nikon D5500 camera on a tripod, white led light connected to a rheostat, and sample stand, all enclosed in a box to occlude ambient lights. The box was painted flat black, and dark fabric hung inside to absorb stray light. A photodiode sensor was fitted inside the box to monitor the intensity of the white led light. The camera inbuilt flash was disabled as the led light serves as the only lighting source. A rheostat attached to the led light enabled us capture images at different intensities (0.7I-I). Capturing the soil images at these different light exposures offers the possibility of replicating the different shades of

Table 1
Soil classification as labeled by the experts.

Sample	Sand (%)	Clay + Silt (%)	Expert's Classification
A	71	29	Suitable
B	78	22	Suitable
C	65	35	Suitable
D	43	57	Unsuitable
E	54	46	Unsuitable
F	42	58	Unsuitable

the soil sample's color due to moisture content. As it has been shown [20,50,51], the moisture content of soil will affect the refractive index and consequently, the tints and shades of its color. Even though all the laterites samples were air-dried for same time interval before capturing, we seek to test the effect of soil's moisture on our technique.

Images were captured at full HD resolution and cropped to 400 × 400-pixel images. Three datasets were populated with the captured images as follows. The training dataset (database) was populated with 1050 images of soil sample A. The test dataset contains 50 images of soil selected randomly from the training dataset, and 50 captured laterite sample D (in Table 1). The validation (here also test) dataset has 50 images of all the soil sample A-F. All images used in the experiment were captured in raw format remotely by 'Wireless Mobile Utility' mobile app. Capturing images in raw format retains the pristine color and information as captured by the camera's sensor.

4.3. Image analysis

The digital camera stores images in RGB (Red, Green, and Blue) color space. While this color space is easy to implement and computationally less expensive for image processing algorithms, the mixture of the chrominance (Color related information) with the luminance (intensity related data) makes it unsuitable under different illumination or different color shades and tints. On the other hand, HSV (Hue, Saturation, and Value) color space represents the chrominance information (Hue) in a separate channel from the luminance (Saturation and Value) [52]. Hence, the HSV color space enables us decouple the effects of luminance variation and soil moisture on the laterites soil images. Fig. 6 shows the effect of a laterite image in HSV color space under different illumination in a 3D histogram. It can be observed that the density plot of the images in HSV color space, remains invariant even though the intensity of the images change (similar trend was observed for other laterite samples not shown). Another advantage of using the HSV over the RGB and other colorspaces is demonstrated in Fig. 5. From the figure, it can be easily observed that the image's pixels in the RGB and LAB color spaces are not confined to a single channel.

On the other hand, the pixels are seen to be more confined in the HSV colorspace. Also, Fig. 7 shows the images of two different laterite images under same illumination in the HSV color space. From inspection, one can quickly establish a difference based on the 3D-density plot. This difference was not discernible in other color spaces (not shown). Given these behaviors of the HSV color space for the images, all the images were transformed from the RGB to HSV colorspace.

4.4. Color histogram

The simplicity of the color histogram is time efficient computationally in terms of image retrieval and has been shown to be robust, computationally efficient, and to effectively represent the color content of an image [53]. Given these facts, the histogram was deployed to extract the samples' color. The color histogram characterizes an image by quantizing the colors within the image and counting the number of pixels of each color [54]. The color features were afterwards stored in the form of real-valued multi-dimensional vectors as follows.

$$\boldsymbol{mj} = \{\boldsymbol{mj}[1], \boldsymbol{mj}[2], \boldsymbol{mj}[3], \ldots, \boldsymbol{mj}[\boldsymbol{i}], \ldots, \boldsymbol{mj}[\boldsymbol{n}]\} \quad (1)$$

where $\boldsymbol{mj}$ is the vector representing an image of the j element in a given dataset, i is the color bin in the color histogram, $\boldsymbol{mj}[\boldsymbol{i}]$, is the number of pixels of color i on the image and n total number of bins used in the color histogram. In this way, once the color histograms of the images have been created after image processing steps described in preceding session, a similarity measure (define in eq. (4)) can be used to quantify the similarity between a query image and those in the database.

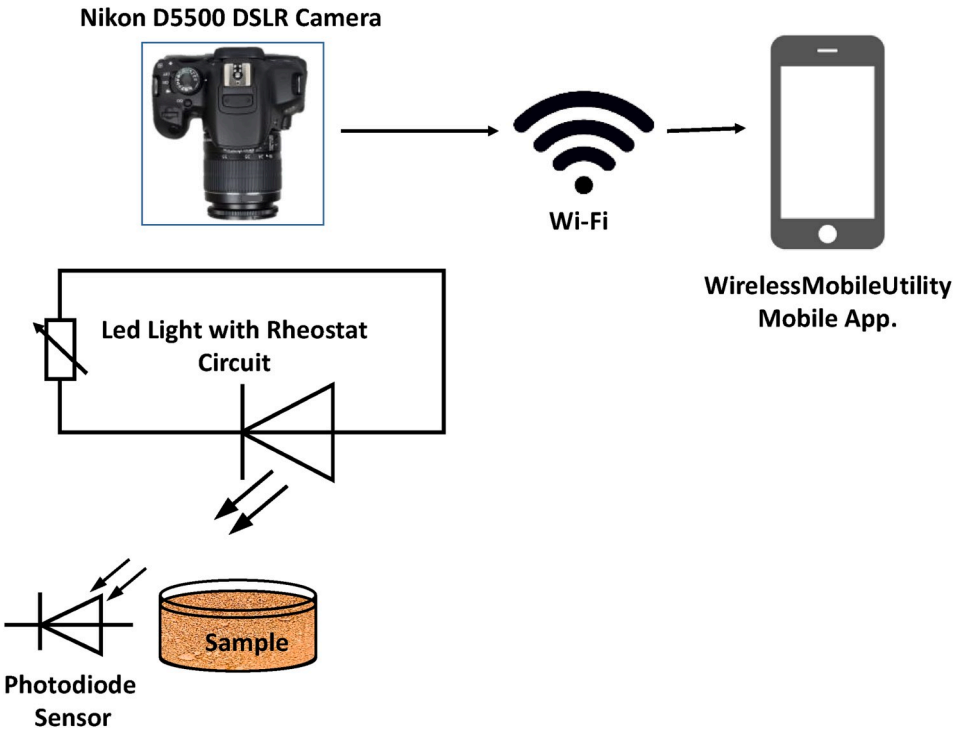

Fig. 3. In-house fabricated capturing system.

Moreover, a critical parameter in color histogram is binning. Binning is a way of grouping the pixels in the captured images into categories. The bin ensures that only a sizeable amount of data is compared across images rather than comparing pixel to pixel. The bin size in a histogram is directly proportional to the discriminatory power of the histogram [55].Table 2 shows the result of different bin sizes on the similarity measure (using cosine similarity measurement) of laterite soil (sample A) as a query image against a database containing another instance of itself and two non-laterites soil (100% sand). From the table, it can be observed that while the discriminatory power of the model improves (ability to distinguish differences between the query image and the database images correctly) with increasing bin, the model performance wanes (similarity of the physical object against its image is less than ideal). Although some, as proposed in the literature, suggested techniques for selecting the bin size for generating color histogram, there is no consensus methodology [56]. Some of the proposed methods include using a clustering method to find the K best colors to select the bin size. It has also been suggested that using the bins with the most substantial pixel number as the descriptor of the histogram will suffice [55]. Some authors suggested using 5–20 bins for real dataset histogram [56]. The last suggestion seems to agree with our results in Table 2. In this work, a bin size of eight was used for the color histogram.

In addition to optimizing or reducing processing time, we observed that selecting the optimal bin size also streamlines the classification technique by rendering the image segmentation process unnecessary. For instance, Fig. 4 shows one of the images of laterite captured under the setup box we referred to in previous session. Image segmentation is usually performed to mask out the shadow effects around the soil particles. Such effects would have been more pronounced at higher bin sizes. However, by selecting the optimal bin size, we can average out the effects of shadows around the soil particles, hence eliminating the need for image segmentation.

5. Methodology

5.1. Training the model

We used a statistical model developed in previous works [48,57] to train our dataset. The model was developed to establish nanoscale compositional heterogeneity from experimental force measurement in atomic force microscopy (AFM). The statistics developed in that work however are universally valid for any task requiring image processing. In the context of this work, we deployed the model as a binary classifier to establish a decision boundary line (threshold) between the suitable laterite soil types and the unsuitable ones after computing the similarity index. For completeness, we explore the basic philosophy underlying the model by to then demonstrate how we used it to train the dataset in this work.

We first define a set $T_n = \{\boldsymbol{m}_1, \boldsymbol{m}_2 \ldots \boldsymbol{m}_n\}$, where the elements are the feature vectors (color histogram as described in section in IV. D) of all the images in the training dataset (database). Let $t_N = \{\boldsymbol{m}_1, \boldsymbol{m}_2 \ldots \boldsymbol{m}_N\}$ be a sample of size N, where $t_N \subseteq T_n$ and N $\leq$ n. Now q can be defined to be the set containing all the unknown image datasets to be classified by assigning binary labels (Suitable & Unsuitable). We can find a decision boundary line D_T in 2D space separating the elements of the set q into appropriate the labels. We define this line quantitatively as

$$\boldsymbol{D_T} = \boldsymbol{s_T} - \boldsymbol{\lambda\sigma}(\boldsymbol{N}) \tag{2}$$

where s_T is the mean cosine similarity index between all the elements of the set t_N , populated with N images of laterite sample T, with an instance of itself, λ ($\lambda > 0$) is a parameter that can be fine-tuned to control the position of the decision boundary line in space and $\sigma(N)$ is the standard deviation of the selected sample N. The mean value s_T can be derived as follows.

From the definition of s_T , the query image (in the test dataset) is selected randomly from T_n and can be represented with a singleton $q = \{\boldsymbol{m_r}\}$, where $\boldsymbol{m_r} \in T_n$.

Therefore, the cosine similarity index between all the images in the dataset t_N plus the singleton q, can be represented as a list of scalar quantities -$Ls(t_N, T)$.

$$\boldsymbol{Ls}(\boldsymbol{t_N}, T) = \{\boldsymbol{c}_{Tt_1}, \boldsymbol{c}_{Tt_2}, \ldots \boldsymbol{c}_{Tt_N}\} \tag{3}$$

where

$$\boldsymbol{c}_{Tt_i} = \frac{\boldsymbol{m_r}.\boldsymbol{m_i}}{\|\boldsymbol{m_r}\|\|\boldsymbol{m_i}\|} \quad (i = 1, 2 \ldots N)$$

Finally, s_T can be defined in the following compact form

$$\boldsymbol{s_T} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{c}_{Tt_i} \quad (1 < \boldsymbol{N} \leq \boldsymbol{n}) \tag{4}$$

The goal of the training stage is to fit the N parameters in eq. (4) and λ in eq. (2) in order to produce the optimal line D_T required, i.e. requiring a minimum amount of data to be collected in order to preform statistics without allowing for inconsistencies between model predictions and real outcome, to classify elements in the query image into appropriate labels. We now seek to provide an example that will further provide an intuitive interpretation to the model.

First we generate a list, with all the elements in training dataset T_n but with two different query images from Table 1 (Sample A and D). In line with the convention adopted in eq. (3), the list will be, $Ls(T_n, A)$ and $Ls(T_n, D)$, i.e., when N=n t_N=T_n and A and D represents elements in the singleton q. The elements in both lists can be represented graphical in Fig. 8. From the plot, the blue circles represent sample A, classified as suitable laterite by the expert, while the orange circles correspond to sample D classified as one of the unsuitable samples. As expected, the value of the blue circles should be closer to 1 than the orange circles (Cosine Similarity of two similar vectors), because all the images in the training dataset T_n are its instances. From Fig. 8, we note that there is no discernible gap between the two-population sets (list) making classification challenging.

While averaging the populations of both lists might be tempting as it might allow us to ensure some form of "theoretically valid" disentanglement between suitable and unsuitable samples, this it is often impractical and computationally costly. A more practical way to proceed relates to looking for a sampling set t_N with reduced size N that is representative of the set T_n . Hence s_k is the mean cosine similarity of query image k with the new set t_N. s_A and s_B will be the mean cosine similarity of the query image with sample A and B respectively against the reduced set t_N. s_T . The latter is the mean cosine similarity of query image T with the instances of its elements in the reduced set t_N, while s_A and s_B on the other hand are the mean cosine similarity indices with all the elements in the reduced set (not necessarily their respective instances). It should be noted that since the training set T_n is populated with sample A, the implication is that for the definition of s_T to hold, the query element T=A and $s_T = s_A$ .However for clarity and the sake of terminology, we will stick to the convention T and s_T for the training stage to emphasize that both averages do not necessary have to be equal.

Table 2
Effects of the color histogram's bin size on the similarity index results.

Bin Size	QUERY	DATABASE		
		Sample A	Red Sand	Yellow Sand
256	Sample A	0.93	0.26	0
128	Sample A	0.94	0.34	0
64	Sample A	0.96	0.47	0.01
32	Sample A	0.98	0.64	0.01
16	Sample A	0.99	0.68	0.01
8	Sample A	1	0.71	0.18
4	Sample A	1	0.84	0.53
2	Sample A	1	1	0.92
1	Sample A	1	1	1

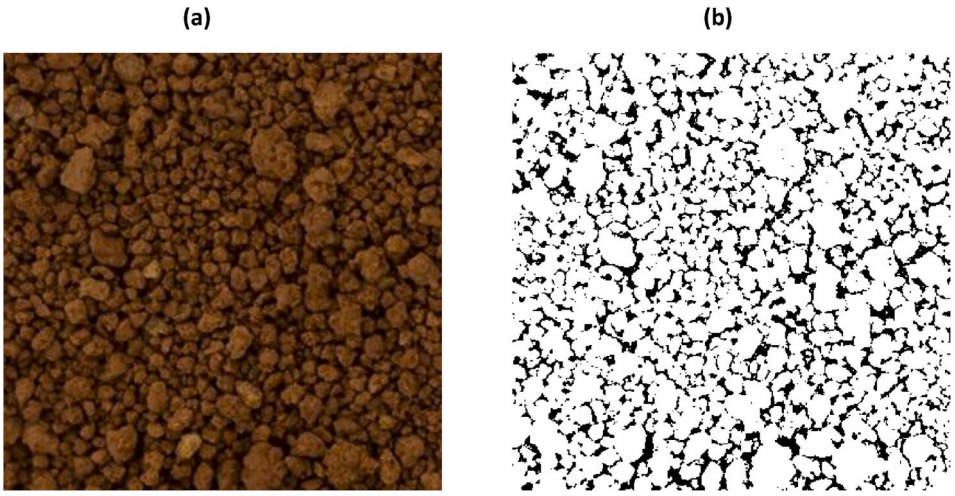

Fig. 4. Captured laterite image without segmentation. (b). Binary mask often used to group pixel value of similar attribute in images (a). Image segmentation can aid the accuracy of classifiers but also add extra layer complication to the model. We circumvent this step with binning.

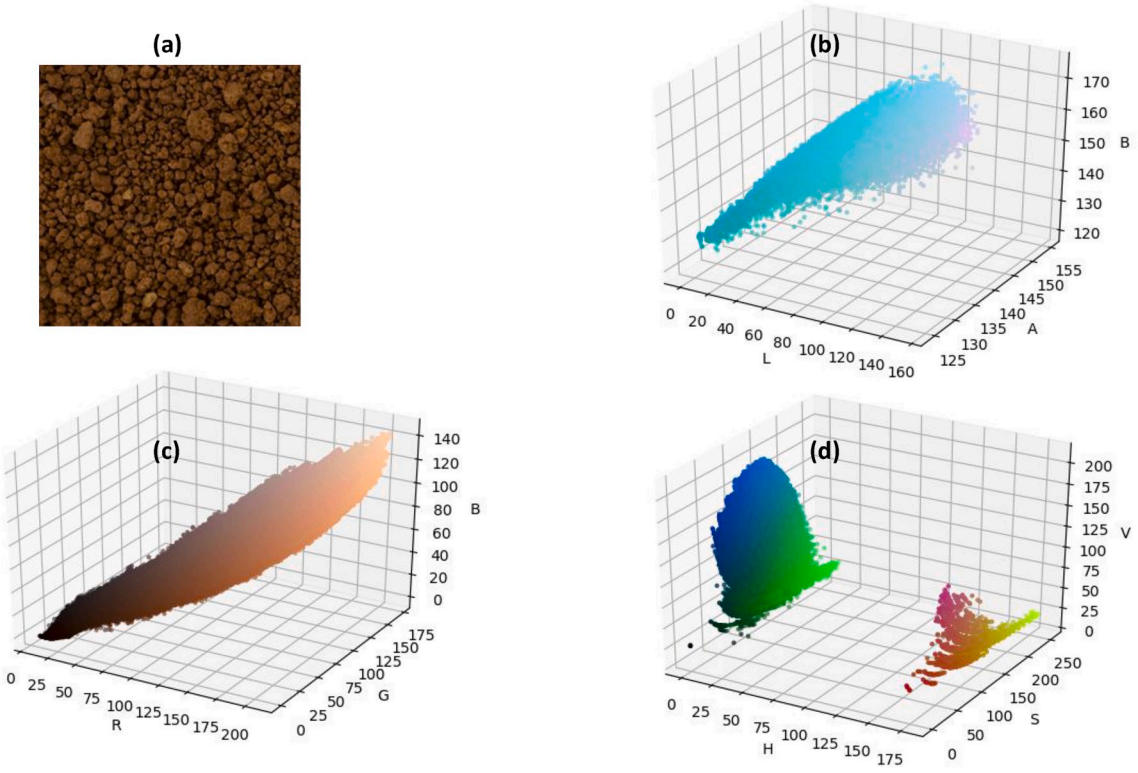

Fig. 5. Laterite soil (a) represented in three different color spaces. (b) - LAB color space. (b) -RGB color space. (c) – HSV color space. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article).

As detailed in Ref. [48], the minimum number N to generate the new set t_N should not be chosen arbitrarily from the population dataset. For instance, we represent the mean for a group of N=50 sampling points with confidence interval (CI) in Fig. 9 using elements in the list $Ls(T_n, T)$. i.e. the elements in the list were grouped into 50 (producing 20 sets of t_N) before the cosine similarity operation with the query image (sample T=A). The inconsistency in the data set is conspicuous, as the CI (when using standard t-distribution) estimates of some set will lie outside the population mean (dash red line). Such inconsistency can lead to erroneous conclusions such as predicting that suitable soil is not suitable and vice versa. Such issues are typically associated to the incapacity of reducing a population to a distribution that can be described via the presuppositions of the Central Limit Theorem of statistics. In short, we need to ensure consistency in such sampling dataset before comparing with a different image. In standard language the problem is this: if we cannot even predict that a given sample of a soil belongs to that soil, comparison between different soils becomes absurd. Having obviated in previous studies that data from soils can be represented and mathematically manipulated with the support of standard statistics that obey the Central Limit Theorem, might have led to inconsistencies.

We show that the number N required by constrains and limits the error interval defined as

$$IE(\lambda) = \lambda\sigma(N) \tag{5}$$

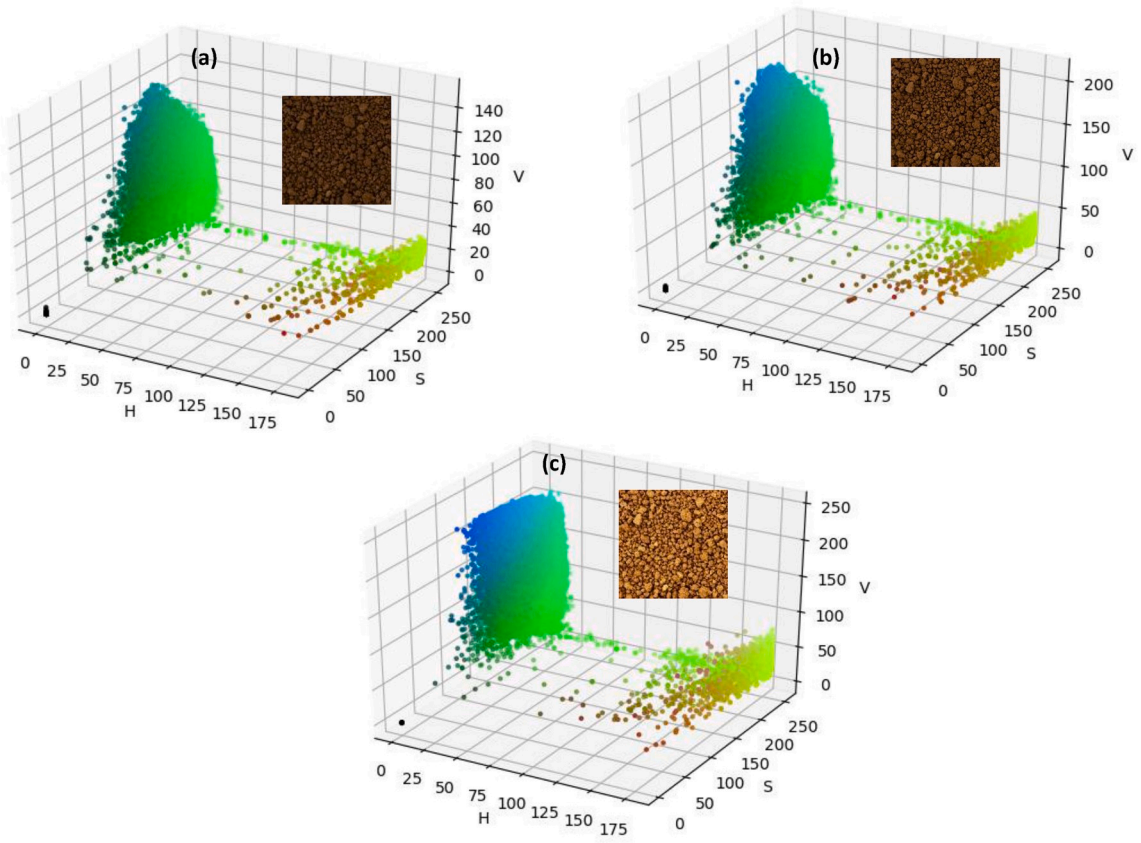

Fig. 6. Laterite soil captured under different light intensity (1.2I -0.7I). (a) - Lower illumination (0.7I). (b) -Normal illumination (I). (c) –Higher illumination (1.2I).

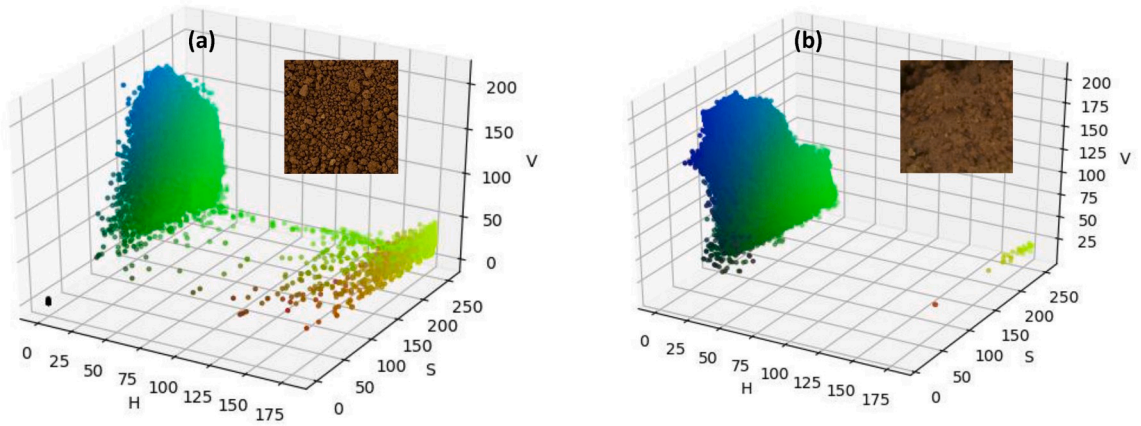

Fig. 7. Two different laterite samples in HSV color space. (a) - Laterite soil labeled "Suitable" by the expert. (b) Laterite soil labeled "Unsuitable" by the expert. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article).

The advantage of this error interval over the standard student t-distribution is that it is generally invariant to the underlying distribution of the dataset and its properties thus avoiding the need of imposing restricting assumptions to the samples or soil.

The parameter λ in equation (5) can be assumed to be constant $\sigma(N) = \sigma$. This assumption is reasonable as evidenced by Fig. 12. Nevertheless, we still experimentally verified this condition as follows. The values of the list $Ls(T_n, T)$ were employed to plot the behavior of the estimates of the standard deviation σ as a function of various N. The value of σ from the plot only increased by 8% from N=2 to N=50 and flattened out afterwards (only $N \leq 200$ shown in the graph). In this regard σ can be considered a constant in the context of eq. (5).

The practical use of eq. (5) now rests on deducing the appropriate value of λ that is consistent with the training dataset. For this purpose, we define a metric - the accuracy ratio $AR(\lambda)$ or inclusion interval in terms of $IE(\lambda)$ as

$$\boldsymbol{AR}(\boldsymbol{\lambda}) = \frac{Excluded(IE(\lambda))}{Total} \tag{6}$$

The term Excluded $(IE\ (\lambda))$ is identified with the number of intervals that exclude the true mean (the mean of the population size –n) whereas Total makes reference to the total number of $IE(\lambda)$*'s* (here n/N). The concept of AR can be illustrated qualitatively by inspecting Fig. 11, where three intervals IE are shown. The first interval on the left does not

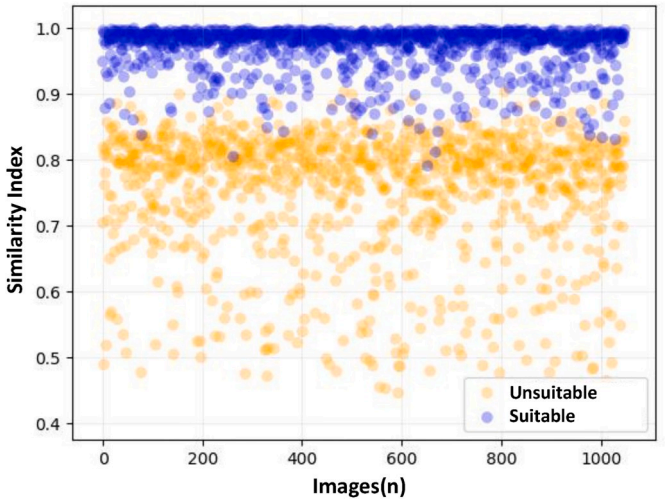

Fig. 8. Similarity index between "Suitable" and "Unsuitable" laterites with the training dataset.

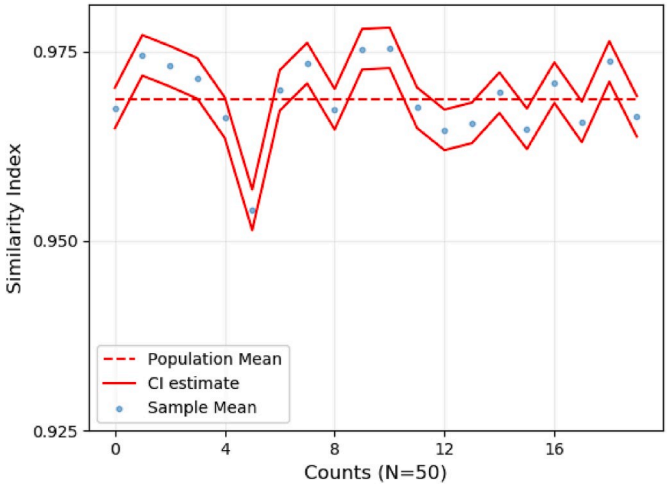

Fig. 9. CI (Red Lines) estimates with mean values (Blue Circles) of samples, 50 data-points each. The interval lies outside the population (dash red lines) for some samples, implying that we cannot disentangle or reliably label "suitable" and "unsuitable" with standard statistical assumptions. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article).

include the true mean (dashed lines) as indicated by the cross while the remaining ones on the right include the mean (tick). In this case, using eq. (6), the resulting AR $= 1/3 \approx 0.33$, indicating a confidence, or accuracy ratio, of 67%. On the other hand, using similar deductions, the AR for Fig. 11 (b) will be AR $= 2/3 \approx 0.66$ and the confidence will be 33%. Hence, we can conclude that for the same precision (IE), the measurement given in Fig. 11, (a) present lower AR (higher accuracy) compared to (b).

The dependence of $AR(\lambda)$ on N, using the training dataset (i.e. the list-$Ls(T_n, T)$ is shown in Fig. 10. The vertical axis is $AR(\lambda)$ and the horizontal axis stands for N while the values of λ range from 0.1 to 0.5 for the plots. From the figure, for a given λ , it can be deduced that $AR(\lambda)$ monotonically decreases with increasing N. The implication of this behavior is that, for a particular λ, there is a minimum N that will comply with the constraint of AR defined in eq. (6). This implies that λ and N cannot be selected arbitrarily. In this work, we seek a confidence interval of at least 95%, i.e. $AR(\lambda) < 0.05$, in our measurements. This roughly implies that the results of our model would agree 95% of the times with the expert. According to the plot in Fig. 10 (a), the minimum number data points required to achieve this level of accuracy is about N $=$ 20–30 with a sample's size at $\lambda = 0.5$ (Magenta color line). This means that higher precision, for a constant accuracy $AR(\lambda) < 0.05$, requires a larger sample size. For example, when λ=0.1 (Blue line) the sample's size required is N=130 in order to achieve an accuracy of 95% ($AR(\lambda) <$ 0.05).

The above concludes the training and statistical treatment of our dataset. We can now proceed to select the parameters for eq. (2) based on the graph in Fig. 10(a). In this work, we used λ=0.5 and N=50 (this corresponded to a confidence interval of 98%) to compute D_T

$$D_T = s_T - 0.5\sigma \tag{7}$$

The practical aspects of our proposed method are (1) we obtain the minimum sample size N (in this case two orders of magnitude smaller than the training dataset) required to establish the difference between two datasets. Knowledge of this parameter saves computational resources. (2) We ensure self-consistency within a given dataset, i.e. we can now randomly select the predicted sampling size N from the training set to represent our population dataset without any concerns regarding the underlying distribution. (3) The method also provides the minimum resolution required ($(0.5\sigma$ in our case) to establish heterogeneity between two datasets. i.e. in reference to our training dataset, given two different test datasets with mean s_A and s_D with same sample size (N) and λ , heterogeneity can only be established between the datasets if $\|s_A$ - $s_D\| \geq 0.5\sigma$.

We can generalize statement (3) above and show that the minimum difference (D_m) between the mean of two datasets accounting for the error can be expressed by

$$D_m = \Delta\mu - 2\lambda\sigma^* \tag{8}$$

where $\Delta\mu = \|s_A$ - $s_D\|$ and σ^* is the mean of the two standard deviation values (See Fig. 11 (c)). From the figure, intuitively, we can state that the constraint to determine a difference between two datasets is

$$D_m > 0 \tag{9}$$

The advantage of eqs. (8) and (9) is that we can now estimate the minimum value of λ necessary to establish heterogeneity between the datasets systematically, from the means and standard deviations as

$$\frac{\Delta\mu}{2\sigma^*} > \lambda \tag{10}$$

From which a maximum (or critical) value λ_c is obtained

$$\lambda_c = \frac{\Delta\mu}{2\sigma^*} \tag{11}$$

In summary, if two datasets are significantly different (images of laterites) for a given sample size N and accuracy $AR(\lambda)$, the value of λ required to determine the decision boundary line will lie in the range

$$0 < \lambda < \lambda_c \tag{12}$$

Finally, we can state the classification conditions for the images to be assigned a label of suitability or unsuitability as follows

$$CEB(s_j, N) = \begin{cases} suitable, & s_j > D_T \\ unsutable, & s_j \leq D_T \end{cases} \tag{13}$$

where s_j is computed for the individual images in the test data and is derived as defined in eq. (4) and where D_T is given by eq. (7) for same sample size N. Fig. 12 (b)–(d) show the result of using eq. (13) with the training dataset-$Ls(T_n, A)$ /$Ls(T_n, T)$, used earlier in Fig. 9. The data were grouped in a set where N=50 ($AR(\lambda) < 0.02$ and λ:$\lambda = \lambda_c = 1.1, \lambda = 0.3$ and $= 1.2$).

We can now randomly select the sample size predicted by the model for comparison since they faithfully contain the population interval. While lower λ values can aid in terms of precision, accuracy is then

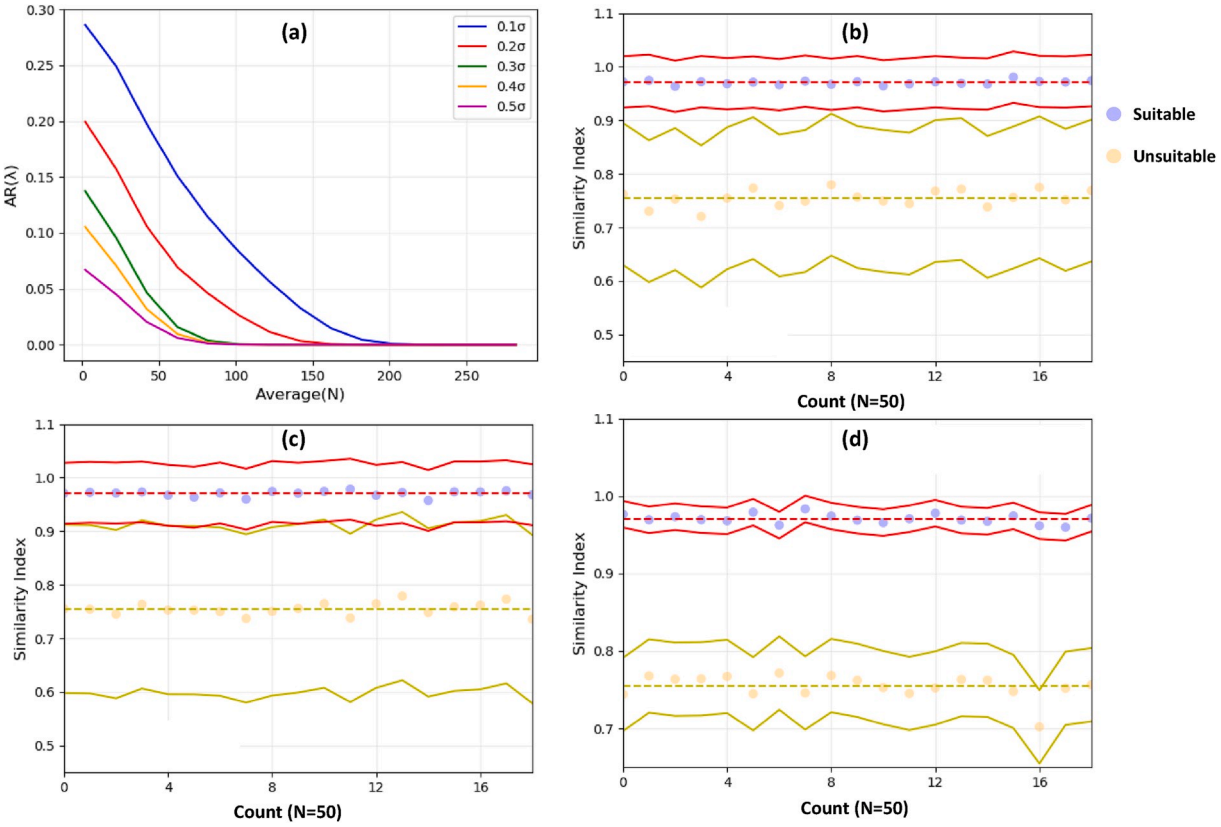

Fig. 10. (a) Behavior of $AR(\lambda)$ on N at different λ (0.1–0.5). Behavior of $IE(\lambda)$'s at N=50 under different λ (b) $\lambda = \lambda_c = 1.1$ (c) $\lambda = 1.2$ (d) $\lambda = 0.3$.

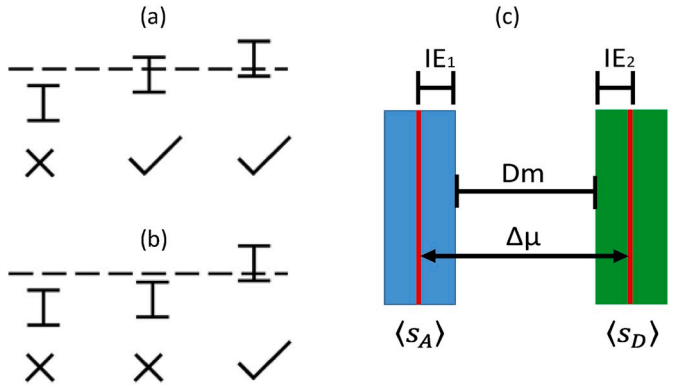

Fig. 11. (a) & (b) Illustration exemplifying the method to numerically compute the accuracy of the interval in the measurement$AR(\lambda)$. $AR(\lambda) = 0.33$in **a** and 0.66 in **b**. (c) Illustration of Dm in relation to precision and averages of the measurements.

compromised. Experimentally we observed that at $\lambda < 0.3$ (not shown), about 15% of the intervals excluded the "true" averages.

6. Results and discussion

6.1. Testing and validation

To test and validate the algorithm, a validation (and simultaneous test since we already got the optimum model) dataset comprising a set of 50 images from all the soil types (Table 1) was used. We further used the derived parameters ($\lambda = 0.5$, and $AR(\lambda) < 0.02$) in the training process to classify the objects in the new validation datasets. That is, we employed our labeled model. According to the constraints of the selected parameters, the minimum number of N required for classification is 50. We trimmed the image in the training datasets to the required N value. For completeness and to emphasize the success of our approach we also show the behavior of selecting different N values: N=10 (below the predicted N value) and N=150 (above the predicted value) keeping λ and $AR(\lambda)$ constant. The decision boundary line as predicted by eq. (7) was used with expression in eq. (13) to assign labels. The results of the comparison are shown in Fig. 12. The F-score was used to quantify the performance. Via Precision and Recall [58]. Precision is the ratio between true positives and predicted positives and Recall is the ratio between true positives and predicted positives. The F-score parameter combines Precision and Recall as

$$F - score = \frac{Precision\ \ X\ \ Recall}{Precision + Recall} \tag{14}$$

The F-score has an advantage over Precision or Recall as a figure of merit since it combines both the specificity (precision) and sensitivity (recall) of both parameters. That is high values in F-score will be obtained if and only if both Recall and Precision are high simultaneously.

Fig. 12 demonstrates our approach achieves to manage the predictions by the expert with an F-score = 0.98 for N=30. The figure also shows how the score falls below 0.95 and above 0.97 for N=10 and N=150 respectively. For completeness, we will like to acknowledge the following as possible limitation of our technique: 1) The F-Score value or predictive power of the algorithm might reduce in the presence of over-damped or humid laterite soil sample, 2) This technique might not be suitable for non-linear classification problem and finally, 3) The proper presentation of data or the quality and resolution of the images in our case is crucial to obtain better F-Score.

In any case, we achieved the main target, namely, to conclusively improve predictive power by simply increasing the number of data points N in out sample for a model that is shown to successfully and routinely label soil as suitable and unsuitable by exploiting image

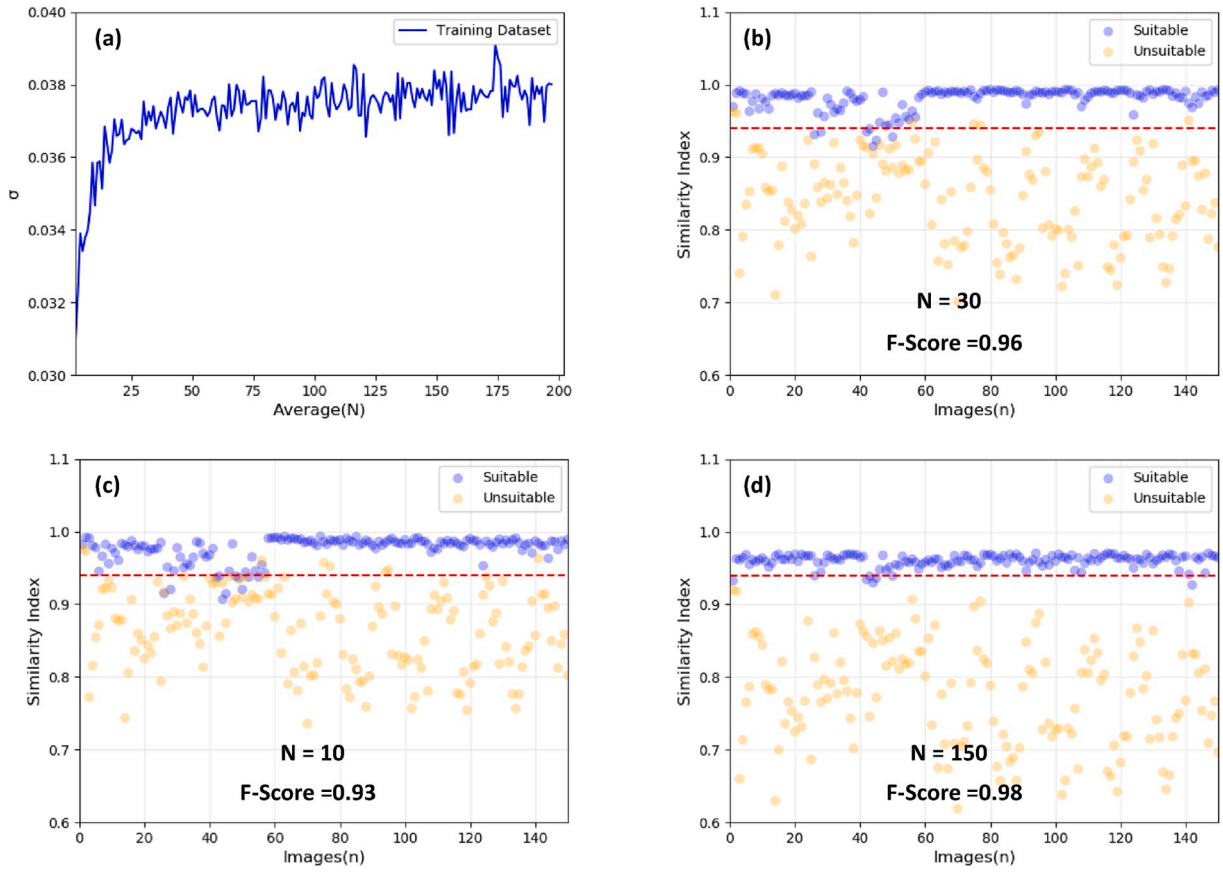

Fig. 12. (a) σ slightly increased with N for the dataset in the training set. (b)–(d) Classification result and F-score in the validation dataset at different N images in the dataset.

processing and statistics only. Since the statistical model is computationally cheap, we expect it can be exploited directly via smart phones in the future.

7. Conclusion

The housing problem in the developing world is concerning because of the lack of "de facto" investment in the development of methods, research of materials and infrastructure and other similar technicalities, but also because of the lack of research resources allocated to solving the problem practically rather than conceptually only. The target is to find a building material that simultaneously helps us to decrease cost, that the people in these countries trust and that does not result in complex procedures to be followed by the natives or in the requirement to highly invest in infrastructure and transport of material. We believe with our approach we have alleviated the problem in all ambits. Technically, we have managed to exploit the expert knowledge of natives by reproducing it via a cheap and robust statistical method that relies on image acquisition only. With the increasing availability of smart phones worldwide, and particularly in the developing countries, our approach could be implemented in a way that the method of soil selection would be available to most natives. In short, the simplicity and efficiency of our technique makes it feasible to implement and distribute in low-cost computing devices, like mobile phones, a robust method that people can trust and the accessibility of which would not be confined to a few experts only.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Tuza A. Olukan: Conceptualization, Methodology, Data curation, Writing - original draft, Software, Investigation, Formal analysis. **Yu-Cheng Chiou:** Software, Validation, Formal analysis. **Cheng Hsiang Chiu:** Software, Visualization, Investigation. **Chia-Yun Lai:** Writing - review & editing, Visualization. **Sergio Santos:** Writing - review & editing, Data curation, Software. **Matteo Chiesa:** Writing - review & editing, Resources, Project administration.

References

[1] R. Bahar, M. Benazzoug, S. Kenai, Performance of compacted cement-stabilised soil, Cement Concr. Compos. 26 (7) (2004) 811–820.
[2] R. Stulz, K. Mukerji, Appropriate building materials: a catalogue of potential solutions, Swiss Centre Appropr. Technol. Skat Publ. Switz. (1988).
[3] C. Oyelami, J.L. Van Rooy, A review of the use of lateritic soils in the construction/development of sustainable housing in Africa: a geological perspective, J. Afr. Earth Sci. 119 (2016) 226–237.
[4] P.N. Lemougna, et al., Laterite based stabilized products for sustainable building applications in tropical countries: review and prospects for the case of Cameroon, Sustainability 3 (1) (2011) 293–305.
[5] E.R. Tuncer, Engineering Behavior and Classification of Lateritic Soils in Relation to Soil Genesis., 1976.
[6] E.R. Tuncer, Engineering Behavior and Classification of Lateritic Soils in Relation to Soil Genesis, 1977.
[7] E. Tuncer, R. Lohnes, An engineering classification for certain basalt-derived lateritic soils, Eng. Geol. 11 (4) (1977) 319–339.
[8] Z.-C. Moh, M.F. Mazhar, Effects of method of preparation on index properties of lateritic soils, in: Soil Mech & Fdn Eng Conf Proc/Mexico/, 1900.
[9] A. Little, The engineering classification of residual torpical soils, in: Soil Mech & Fdn Eng Conf Proc/Mexico/, 1969.

[10] R. Lohnes, R. Fish, T. Demirel, Geotechnical properties of selected Puerto Rican soils in relation to climate and parent rock, Geol. Soc. Am. Bull. 82 (9) (1971) 2617–2624.
[11] D.E. Foote, Soil survey of islands of Kauai, Oahu, Maui, Molokai, and Lanai, State of Hawaii 1, US Government Printing Office, 1972.
[12] M. Gidigasu, The importance of soil genesis in the engineering classification of Ghana soils, Eng. Geol. 5 (2) (1971) 117–161.
[13] D. Moye, Engineering geology for the snow mountain schema, J. Inst. Eng. Aust. 27 (1955) 281–299.
[14] R.O. Fish, Shear Strength and Related Engineering Properties of Selected Puerto Rican Oxisols and Ultisols, Iowa State University, 1971.
[15] E. Ruddock, Properties and position in lateritic ground: some statistical relationships, Soil Mech. Fdn Eng. Conf. Proc. /Mexico/ (1969).
[16] M. Pinard, D.F. Netterberg, D.P. Paige-Green, Review of Specifications for the Use of Laterite in Road Pavements. Final Report of Contract AFCAP/GEN/124: Association of Southern Africa National Road Agency, UK Department of International Development, 2014.
[17] C. CRATerre-EAG, Compressed earth blocks: standards–Technology series No. 11, Brussels: CDI (1998).
[18] E. Adam, A. Agib, Compressed Stabilised Earth Block Manufacture in Sudan, Printed by Graphoprint for UNESCO, France, Paris, 2001.
[19] H. Gunal, et al., Use of chromameter-measured color Parameters in estimating color-related soil variables, Commun. Soil Sci. Plant Anal. 39 (5-6) (2008) 726–740.
[20] C. Humphrey Jr., M. O'Driscoll, Evaluation of soil colors as indicators of the seasonal high water table in coastal North Carolina, Int. J. Soil Sci. 6 (2) (2011) 103.
[21] M. Aitkenhead, et al., Estimating soil properties with a mobile phone, in: Digital Soil Morphometrics, Springer, 2016, pp. 89–110.
[22] S.-O. Chung, et al., Soil texture classification algorithm using RGB characteristics of soil images, IFAC Proc. Vol. 43 (26) (2010) 34–38.
[23] L. Gómez-Robledo, et al., Using the mobile phone as Munsell soil-colour sensor: an experiment under controlled illumination conditions, Comput. Electron. Agric. 99 (2013) 200–208.
[24] M. Aitkenhead, et al., E-smart: environmental sensing for monitoring and advising in real-time, in: International Symposium on Environmental Software Systems, Springer, 2013.
[25] S. Ibáñez-Asensio, et al., Statistical relationships between soil colour and soil attributes in semiarid areas, Biosyst. Eng. 116 (2) (2013) 120–129.
[26] P. Breul, R. Gourves, In field soil characterization: approach based on texture image analysis, J. Geotech. Geoenviron. Eng. 132 (1) (2006) 102–107.
[27] D. Lim, Development of a fruit sorting system using statistical image processing, Korean J. Appl. statistics 16 (1) (2003) 129–140.
[28] F. Pedregosa, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (Oct) (2011) 2825–2830.
[29] S. Raschka, V. Mirjalili, Python Machine Learning, Packt Publishing Ltd, 2017.
[30] G.C. Liles, et al., Developing predictive soil C models for soils using quantitative color measurements, Soil Sci. Soc. Am. J. 77 (6) (2013) 2173–2181.
[31] V.R. Raphael, W. Christian, Towards a quantitative assessment of soil organic carbon using proximally sensed digital imagery, in: 17. World Congress of Soil Science, Bangkok (Thailand), 14-21 Aug 2002, 2002.
[32] R.V. Rossel, et al., Assessment of two reflectance techniques for the quantification of the within-field spatial variability of soil organic carbon, Precis. Agric. (2003) 697–703.
[33] N. Levin, E. Ben-Dor, A. Singer, A digital camera as a tool to measure colour indices and related properties of sandy soils in semi-arid environments, Int. J. Remote Sens. 26 (24) (2005) 5475–5492.
[34] D. Sena Jr., et al., Fall armyworm damaged maize plant identification using digital images, Biosyst. Eng. 85 (4) (2003) 449–454.
[35] K.L. Yam, S.E. Papadakis, A simple digital imaging method for measuring and analyzing color of food surfaces, J. Food Eng. 61 (1) (2004) 137–142.
[36] D.J. King, Airborne multispectral digital camera and video sensors: a critical review of system designs and applications, Can. J. Remote Sens. 21 (3) (1995) 245–273.
[37] C. Dean, T.A. Warner, J.B. McGraw, Suitability of the DCS460c colour digital camera for quantitative remote sensing analysis of vegetation, ISPRS J. Photogrammetry Remote Sens. 55 (2) (2000) 105–118.
[38] M.A. Webster, J. Mollon, Adaptation and the color statistics of natural images, Vis. Res. 37 (23) (1997) 3283–3298.
[39] X. Zhang, N.H. Younan, R. King, Soil texture classification using wavelet transform and maximum likelihood approach, in: IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No. 03CH37477), IEEE, 2003.
[40] Y. Sun, et al., Gabor wavelet image analysis for soil texture classification. In nondestructive sensing for food safety, quality, and natural resources, Int. Soc. Optics Photonics (2004).
[41] B. Bhattacharya, D.P. Solomatine, Machine learning in soil classification, Neural Netw. 19 (2) (2006) 186–195.
[42] Z. Zhao, et al., Predict soil texture distributions using an artificial neural network model, Comput. Electron. Agric. 65 (1) (2009) 36–48.
[43] W. Wu, et al., A comparison of support vector machines, artificial neural network and classification tree for identifying soil texture classes in southwest China, Comput. Electron. Agric. 144 (2018) 86–93.
[44] X. Zhang, V. Vijayaraj, N.H. Younan, Hyperspectral soil texture classification, in: IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data, 2003, IEEE, 2003.
[45] X. Zhang, N.H. Younan, C.G. O'Hara, Wavelet domain statistical hyperspectral soil texture classification, IEEE Trans. Geosci. Remote Sens. 43 (3) (2005) 615–618.
[46] A.D. Vibhute, et al., Soil type classification and mapping using hyperspectral remote sensing data, in: 2015 International Conference on Man and Machine Interfacing (MAMI), IEEE, 2015.
[47] K. Liakos, et al., Machine learning in agriculture: a review, Sensors 18 (8) (2018) 2674.
[48] Y.-H. Chang, et al., Establishing nanoscale heterogeneity with nanoscale force measurements, J. Phys. Chem. C 119 (32) (2015) 18267–18277.
[49] H. Houben, H. Guillaud, Earthen Architecture: A Comprehensive Guide, Intermediate Technology Development Group, London, UK, 1994.
[50] A.L. Kaleita, L.F. Tian, M.C. Hirschi, Relationship between soil moisture content and soil surface reflectance, Trans. ASAE 48 (5) (2005) 1979–1986.
[51] W. Philpot, Spectral reflectance of wetted soils, in: Proceedings of ASD and IEEE GRS, 2010.
[52] S. Sergyán, Color content-based image classification, in: 5th Slovakian-Hungarian Joint Symposium on Applied Machine Intelligence and Informatics, 2007. Citeseer.
[53] M.J. Swain, D.H. Ballard, Color indexing, Int. J. Comput. Vis. 7 (1) (1991) 11–32.
[54] A.R. Kumar, D. Saravanan, Content based image retrieval using color histogram, Int. J. Comput. Sci. Inf. Technol. 4 (2) (2013) 242–245.
[55] F. Long, H. Zhang, D.D. Feng, Fundamentals of content-based image retrieval, in: Multimedia Information Retrieval and Management, Springer, 2003, pp. 1–26.
[56] K. He, G. Meeden, Selecting the number of bins in a histogram: a decision theoretic approach, J. Stat. Plan. Inference 61 (1) (1997) 49–59.
[57] Y.-H. Chang, et al., Divergent surface properties of multidimensional sp 2 carbon allotropes: the effect of aging phenomena, Nanotechnology 27 (29) (2016) 295701.
[58] D.M. Powers, Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, J. Mach. Learn. Technol. 2 (2011) 37–63.