International Journal of Computational Science and Engineering

ISSN online: 1742-7193 - ISSN print: 1742-7185
https://www.inderscience.com/ijcse

Canopy centre-based fuzzy-C-means clustering for enhancement of soil fertility prediction

M. Sujatha, C.D. Jaidhar

DOI: 10.1504/IJCSE.2022.10058486

Article History:

Received:	10 April 2022
Last revised:	05 September 2022
Accepted:	29 September 2022
Published online:	25 January 2024

Canopy centre-based fuzzy-C-means clustering for enhancement of soil fertility prediction

M. Sujatha* and C.D. Jaidhar

Department of Information Technology,
National Institute of Technology Karnataka,
Surathkal, Mangalore – 575025,
Karnataka, India
Email: sujatham.197it002@nitk.edu.in
Email: smsujatha23@gmail.com
Email: jadharcd@nitk.edu.in
*Corresponding author

Abstract: For plants to develop, fertile soil is necessary. Estimating soil parameters based on time change is crucial for enhancing soil fertility. Sentinel-2's remote sensing technology produces images that can be used to gauge soil parameters. In this study, values for soil parameters such as electrical conductivity, pH, organic carbon, and nitrogen are derived using Sentinel-2 data. In order to increase the clustering accuracy, this study suggests using Canopy centre-based fuzzy-C-means clustering and comparing it to manual labelling and other clustering techniques such as Canopy, density-based, expectation-maximisation, farthest-first, k-means, and fuzzy-C-means clustering, its usefulness is demonstrated. The proposed clustering achieved the highest clustering accuracy of 78.42%. Machine learning-based classifiers were applied to classify soil fertility, including Naive Bayes, support vector machine, decision trees, and random forest (RF). Dataset labelled with the proposed RF clustering classifier achieves a high classification accuracy of 99.69% with ten-fold cross-validation.

Keywords: clustering; classification; machine learning; remote sensing; soil fertility.

Reference to this paper should be made as follows: Sujatha, M. and Jaidhar, C.D. (2024) 'Canopy centre-based fuzzy-C-means clustering for enhancement of soil fertility prediction', *Int. J. Computational Science and Engineering*, Vol. 27, No. 1, pp.90–102.

Biographical notes: M. Sujatha is currently pursuing a Full-Time PhD in the Department of Information Technology, National Institute of Technology, Karnataka, Surathkal, India. Her research interests include machine learning, deep learning and precision agriculture.

C.D. Jaidhar is presently working as an Associate Professor at the National Institute of Technology Karnataka, Surathkal, India. His primary research area includes computer networks, cyber security, internet of things and precision agriculture.

1 Introduction

The most efficient use of resources and improved management of agriculture are made possible by understanding the variation in soil. In order to increase agriculture productivity, plant nutrients must be optimised in a sustainable manner (FAO, 2022). For site-specific fertilisation, soil maps do not offer sufficient qualitative information. Therefore, quantitative evaluation of significant soil parameters utilising laboratory tests, proximity sensors, airborne sensors, or remote sensing information is crucial (Gholizadeh et al., 2018). However, laboratory analysis produces contaminants and is also expensive. The high maintenance and labor requirements of hyperspectral aerial sensors increase the cost of soil analysis.

In several domains, including cloud computing (Li et al., 2017), disaster management (Chen et al., 2020), and agricultural yield prediction (Divakar et al., 2022), remote sensing is used. It is possible to provide cost-effective and timely spatial-temporal information on the soil using remote sensing as a key data source. For instance, land change and vegetation detection are examples where remote sensing has been utilised to support decision-making (Tayeb and Fizazi (2020)). Additionally, the researchers employ remote sensing methods with Sentinel 1-2, Landsat 8 OLI, EnMAP, and HyspIRI to estimate soil parameters (Yang and Guo (2019)). Sentinel-2 offers high-resolution optical soil images that can be utilised to measure different parameters of soil, such as electrical conductivity (*EC*), *pH*, organic carbon (*OC*), and nitrogen (*N*).

The process of grouping data points based on similarity is known as clustering (Shahmoradi and Lee, 2022). Without labelling the data, it uses the similarity or distance between data points to group them (Kim et al., 2020). Different clustering techniques were utilised in this research work to label the dataset automatically. The various clustering methods, including Canopy, density-based, expectation-maximisation, farthest-first, k-means, and fuzzy-C-means (FCMs), can be used to classify the dataset's data points and, thus, to label it. Several fields, including the cloud computing environment (Samandi and Mukhopadhyay, 2021) and the detection of intrusions (Ma and Li, 2020 have employed machine learning-based classifications. It has recently been widely employed in agriculture for various purposes, including crop yield prediction, pest monitoring, water management, crop disease monitoring, and the management of fertiliser, etc. (Priya and Ramesh, 2020). We employed machine learning-based classifiers, including Naive Bayes (NB), support vector machines (SVM), decision tree (DT), and random forest (RF), to classify soil fertility and measured their classification performance using clustered datasets.

This research work has made the following significant contributions:

- Derived the values of soil parameters such *EC*, *pH*, *OC* and *N* using Sentinel-2 spectral data.
- The resulting dataset is labelled using a variety of clustering algorithms, and the performance of the clustering approaches is compared.
- Canopy centre-based FCMs clustering is proposed to increase clustering and classification accuracy.
- The performance of the proposed approach is assessed through experiments.

The structure of this paper is as follows: Section 2 reviews previous research on the measurement of soil fertility and discusses the clustering approach utilised in various fields. The geographic study area and research methodology are described in Section 3. A discussion on the estimation of soil parameters using Sentinel-2 spectral bands is provided in Section 4. The various clustering approaches employed in this work are discussed in Section 5, along with comparisons. In Section 6, the proposed clustering approach is described. Section 7 provides conclusions and suggestions for further research.

2 Related work

Remote sensing is a promising field of research, combining several research domains, including weather forecasts, land cover changes, natural disaster forecasts, and satellite images. Many artificial intelligence methods have been employed in these fields to increase accuracy and improve data analysis. Tayeb and Fizazi (2020) used extractor-MLP and SVM to classify six physical soil parameters from a Statlog Landsat remote sensing dataset, including red soil, cotton crop, grey soil, damp grey soil, soil with vegetation stubble, and very damp grey soil. Ye et al. (2018) found that remotely sensed soil data, which includes parameters like *pH*, soil temperature, and humidity, enhances irrigation to boost agricultural output.

Many researchers have classified soil fertility based on soil chemical parameters using machine learning-based classifiers. To classify soil fertility, the authors employed laboratory-measured soil parameters. Sirsat et al. (2017) employed bagging, AdaBoost, extreme learning machine, RF, and SVM to classify soil based on laboratory measures of *EC*, *pH*, iron (*Fe*), manganese (*Mn*), potassium oxide, phosphorus pentoxide, nitrous oxide, sulphate, and zinc (*Zn*) and also soil type. Wang et al. (2018) estimated soil *OC* stocks using ground measurements of *OC* stocks, utilising RF and boosted regression tree classifier. Utilising laboratory measurements of *B*, *Cu*, *EC*, nitrous oxide, potassium oxide, sulfate, and *Zn*, as well as village-wise fertility indices of *Fe*, *OC*, *Mn*, and phosphorus pentoxide, Sirsat et al. (2018) developed a soil fertility model using extreme learning machine, bagging, boosting, and extremely randomised regression. Chougule et al. (2019) used k-means clustering to label the dataset and an RF classifier to classify the soil fertility using laboratory-measured soil data comprising of *N*, *P*, and potassium (*K*). Dasgupta et al. (2022) used ground measurements of *K*, *Ca*, magnesium (*Mg*), *Fe*, *Cu*, *Zn*, *Mn*, *B*, *K/Mg* ratio, total exchangeable bases, and sulphur availability index to predict soil fertility via RF, support vector regression, stepwise multiple linear regression.

Gholizadeh et al. (2018) investigated the capability of classifying soil fertility using data collected from remote sensing and laboratory measurements. With the spatiotemporal data estimated using remotely sensed spectral data, soil fertility can be classified precisely and accurately (Gholizadeh et al., 2018). Gholizadeh et al. (2018) compared the capabilities of Sentinel-2 for mapping soil *OC* and texture with those obtained from hyperspectral sensors and laboratory measurements and developed a site-specific model utilising multivariate regression and boosted regression tree. The study showed that, when compared to non-remote sensing, the remote sensing approach obtained the highest accuracy. Khanal et al. (2018) claims that site-specific remotely sensed soil image maps obtained using Lidar satellite could be used to accurately predict soil parameters such as *pH*, organic matter, *K*, magnesium, cation exchange capacity, and crop yield using machine learning-based classifiers such as gradient boosting, RF, cubist, neural networks, and SVM. Yang and Guo (2019) predicted various soil parameters in coastal wetlands with dense vegetation cover utilising synthetic aperture radar data. Using Landsat-8 OLI and Sentinel-2 visible bands, as well as linear regression and multiple linear regression approaches, Gorji et al. (2020) compared the estimation of soil salinity. The difference between the accuracy obtained for multiple linear regression using Landsat and Sentinel-2 was negligible. Hengl et al. (2021) demonstrated the use of ensemble machine learning to estimate *pH* and *OC* using Landsat and Sentinel

bands. Aksoy et al. (2022) implemented classification and regression trees, RF, and support vector regression to measure the correlation between data derived from Landsat-8 OLI, Sentinel-2, and ground measurements of *EC*.

A review of prior clustering studies in a variety of fields has been conducted. It was found that the cluster centre influences clustering accuracy. Zhang et al. (2018) developed the k-means method based on the density Canopy technique utilising UCI datasets. By choosing the cluster centres using density-based clustering, the researchers noted that clustering accuracy and stability were enhanced. To reduce the rate of false alarms during security events, Wang et al. (2019) developed an improved density peak clustering using a time gap as a threshold. Du et al. (2022) developed a threshold for an automatic density peaks clustering technique to identify the peak density and used the interquartile range and standard deviation to define a number of local densities. Using the UCI dataset, the authors found that the novel method outperformed k-means and FCM in terms of clustering accuracy. Wu et al. (2021) employed FCM to cluster the operation of mineral flotation in gold mines. Zhu et al. (2022) proposed that choosing clusters based on internal validity indices and inter-cluster variance could enhance k-means clustering. As starting cluster numbers, the authors used integers from the range [2, $\sqrt{n}$] (where n is the total number of data points).

A few soil parameters were estimated by the researchers using remote sensing data, and the soil dataset generated was manually labelled. Sentinel-2 is used in this research work to estimate four soil parameters, and the dataset is labelled using a clustering method. This study also suggests Canopy centre-based FCMs clustering, which chooses cluster centres depending on the technique used in the outperforming state-of-the-art clustering approach.

3 Study area and methodology

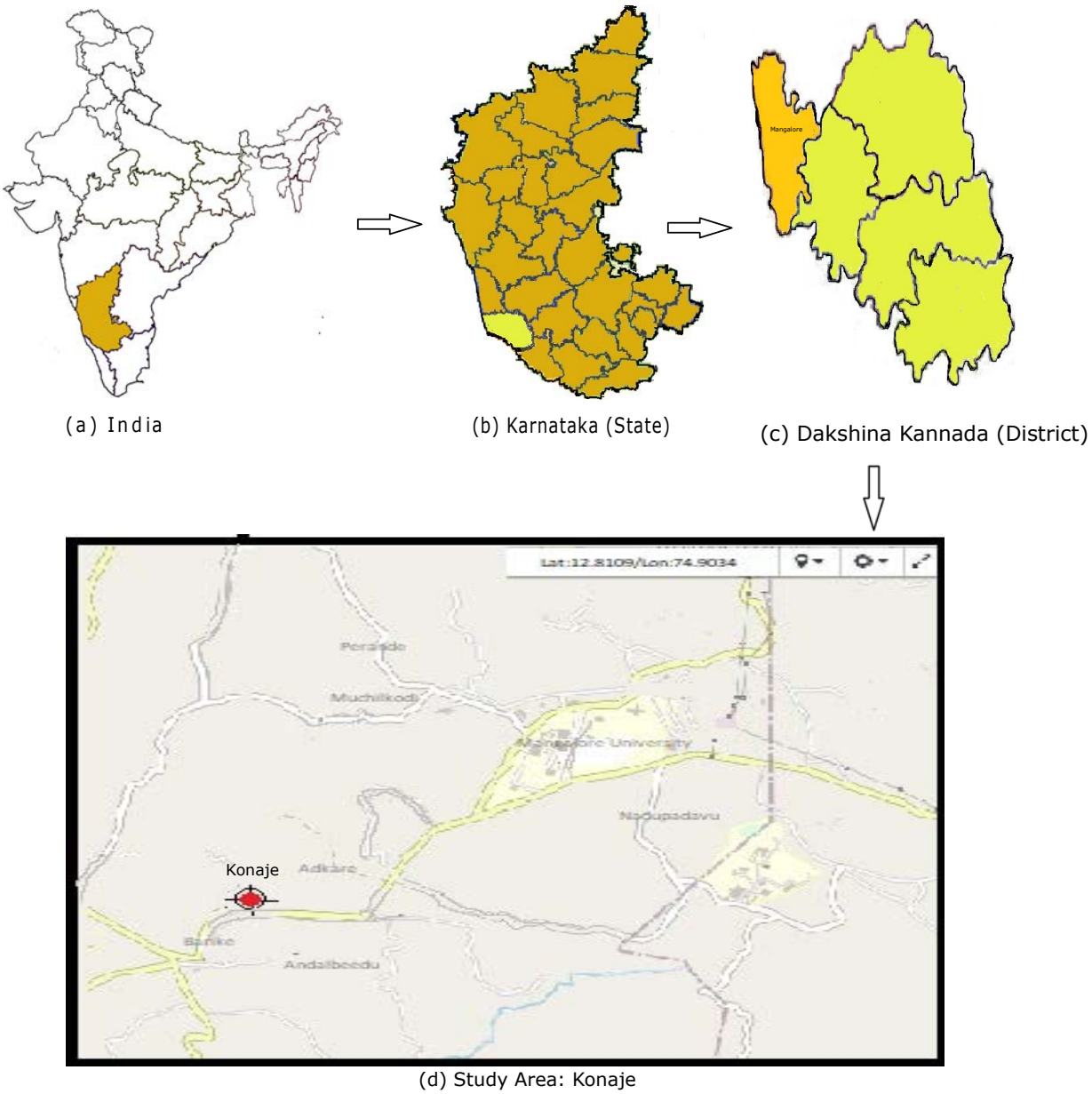
The study region employed in this work is located at Konaje, Mangalore town in Dakshina Kannada (District), Karnataka (State), India, between 12.80593353 latitude and 74.906469 longitude, as shown in Figure 1. Dakshina Kannada has alluvial, laterite, red laterite, sandy loam, and red loamy soil as its main soil types. The region experiences roughly 3,000 mm of yearly rainfall and has a hot, humid environment. Climatic conditions are favourable for the development of acidic soils. The geology of Mangalore is mainly hard laterite in hilly and sandy areas along the Arabian coast, with a humid tropical climate.

Figure 2 presents the methodology used in this work to classify soil fertility. Data from Sentinel-2 (2020) were collected for the chosen study area between 14th November 2015, and 23rd October 2021. With the help of Google Earth Engine code, spectral bands were extracted using the Python programming language. Seven spectral bands, B3, B4, B5, B8, B9, B11, and B12, were employed to get the soil parameters out of the 13 retrieved spectral bands. A combination of spectral bands was used to compute soil parameters, including *EC*, *pH*, *OC*, and *N*. In the spectral band information, bits 10 and 11 stands in for clouds and cirrus clouds, respectively. By implementing a bitwise mask on bits 10 and 11 of the spectral band information, QA60 was utilised to mask clouds. The WEKA tool (WEKA, 2021) was used to perform data preprocessing in order to remove redundant data. The Sentinel-2 dataset is used to group data points using a variety of clustering techniques, including Canopy, density-based, expectation-maximisation, farthest-first, k-means, and FCM clustering. For each clustering algorithm, the number of clusters was fixed to 3. Instances in the dataset were assigned a LOW, MEDIUM, or HIGH label based on the way the data points were clustered. This study evaluated and compared the clustering algorithms' accuracy to manual labelling. Additionally, we used machine learning-based classifiers to classify soil fertility into three categories: LOW, MEDIUM, and HIGH soil fertility. FCM clustering performs poorly because the initial centroids are produced randomly. Therefore, based on the strategy utilised in outperforming clustering algorithms, cluster centroids were produced in this research work. The Canopy centre-based FCMs clustering technique is proposed in this research work to increase clustering and classification accuracy. Figure 3 depicts the steps of the proposed approach. The fuzzy parameter (m) was employed to enhance the fuzzification process. By varying 'm' from 1.1 to 1.9 with a 0.1 increment, the optimal values for 'm' were discovered. The Canopy approach was used to obtain the initial centroids, which were then employed in the membership function. The membership function and cluster centroids were updated iteratively until every data point had been allocated to a cluster. The clusters are also labelled as LOW, MEDIUM, or HIGH.

4 Estimation of soil parameters using Sentinel-2

The European Space Agency's recent release of multispectral Sentinel-2 satellite data offers an innovative method for collecting images with high spatial resolution, 290 km swath width, and high repetition rate. Sentinel-2, a multispectral satellite, offers 13 spectral bands: B1 – aerosol, B2 – blue, B3 – green, B4 – red, B5 – red edge 1, B6 – red edge 2, B7 – red edge 3, B8 – near infra-red, B8A – red edge 4, B9 – water vapour, B10 – cirrus, B11 – short wave infra-red 1, B12 – short wave infra-red 2, with spectral resolution ranging from 20 to 180 nm and spatial resolution ranging from 10 to 60 m, may be used to estimate different parameters. Additionally, it provides three cloud masks QA10, QA20, and QA60, with pixel sizes of 10 metres, 20 metres, and 60 metres, respectively (Vaudour et al., 2019). As illustrated in Figure 4, spectral bands B3, B4, B5, B8, B9, B11, and B12 are utilised to estimate soil parameters along with cloud mask band QA60. The spectral band combination used to calculate soil parameters from Sentinel-2 images is shown in Figure 5.

Figure 1 Study area: village located at the latitude of 12.80593353 and longitude of 74.906469 in Dakshina Kannada (District), Karnataka (State), India (see online version for colours)



In this study, *EC, pH, OC*, and *N* are calculated using spectral bands from Sentinel-2.

Soil *EC* indicates the amount of salinity in soil (NRCS-USDA, 2020) and provides information on nutrient availability and loss, soil texture, and water availability (Al-Gaadi et al., 2021). Fertile soils are those with an EC value of 1.6 mS/cm or less, whereas low fertility soils have an *EC* value of greater than 2.5 mS/cm. Equation (1) was used to get the *EC* values for the spectral bands B3, B4, B8, and B12.

$$EC = 2.5 * \frac{B12}{B8} - 1.6 * \frac{B4}{B8} + 1.6 * \frac{B3}{B8} \tag{1}$$

Soil *pH* assesses the amount of hydrogen ions present in soil solutions (Tharavathy, 2016). A *pH* of 7 in the soil indicates a neutral soil and is considered to be very fertile. Acidic soil is caused by water or moisture in the soil, which lowers *pH* below 6.5. Low *pH* soil is considered to be less fertile. The spectral bands used for *pH* calculation were B4, B9 and B12 (Hengl et al., 2021) as in equation (2).

$$pH = 7 * \frac{B4}{B12} - 6.5 * \frac{B9}{B12} \tag{2}$$

Soil *OC* indicates the amount of OC present in the soil. Soils with *OC* values above 0.75% are considered very fertile. To calculate *OC*, the spectral bands B4, B5, B11, and B12 (Hengl et al., 2021) were used as shown in equation (3).

$$OC = 0.75 * \frac{(B12 - B4)}{B4} + 0.75 * \frac{B5}{B11} + 0.75 * \frac{B12}{B11} - 0.05 \tag{3}$$

Figure 2 Clustering and soil fertility classification using Sentinel-2 data (see online version for colours)

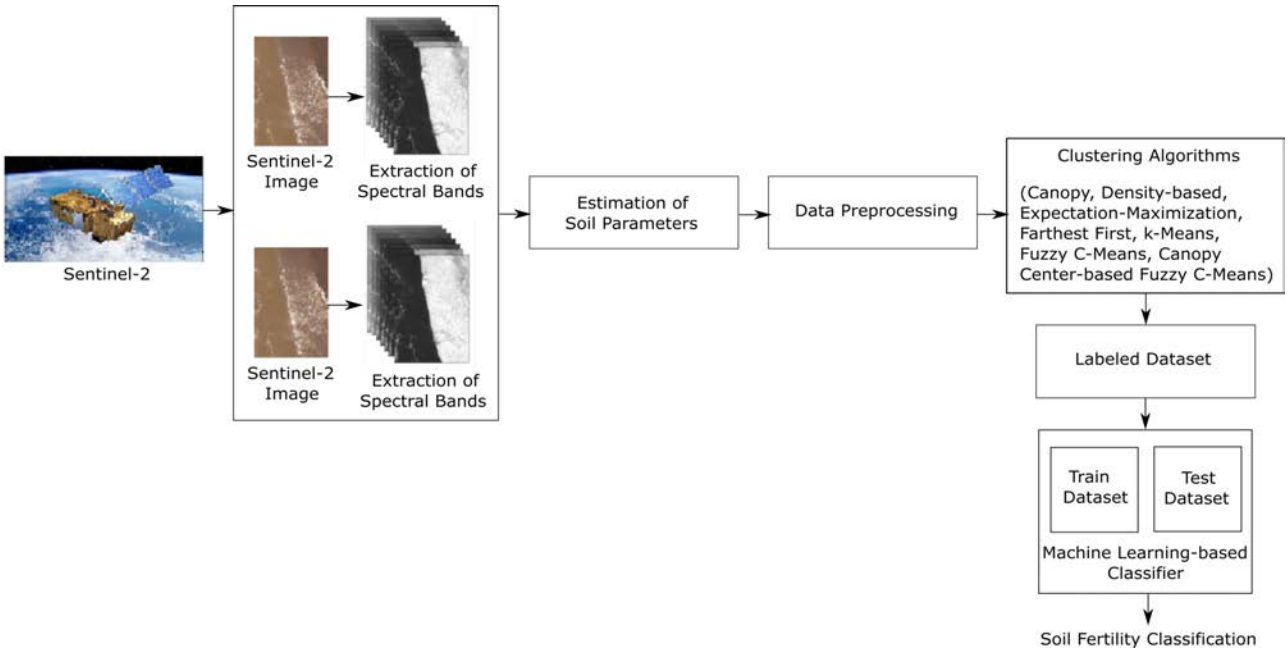
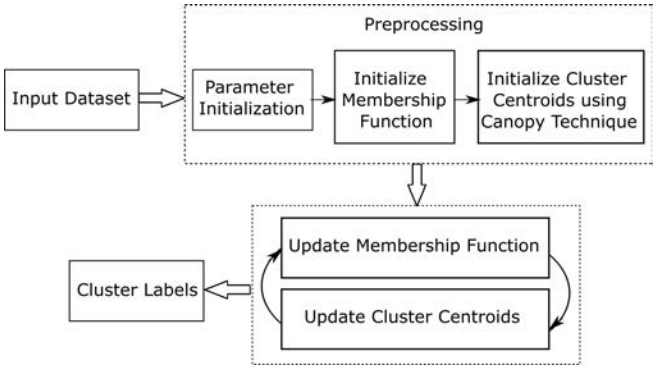


Figure 3 Proposed Canopy centre-based FCMs clustering



Soil N is proportional to the red to green index, B4/B3 (Xu et al., 2018). The higher the N value, the higher the vegetation (Mashaba-Munghemezulu et al., 2021). As a result, we must subtract the normalised differential vegetation index and the normalised red-edge index. Normalised difference vegetation index can be calculated as (B8 – B4)/(B8 + B4) and normalised red-edge index by using (B8 – B5)/(B8 + B5). High fertility is defined as having a nitrogen content of 560 kg/ha or more, while low fertility is defined as having a nitrogen content of less than 280 kg/ha. Thus, as given in equation (4), N is determined using the spectral bands B3, B4, B5, and B8.

$$N = 560 * \frac{B4}{B3} - \frac{280}{100} * \frac{(B8 - B4)}{(B8 + B4)} + \frac{280}{100} * \frac{(B8 - B5)}{(B8 + B5)} \tag{4}$$

The results were compared to soil-health data (Soil-Health Data, 2021), which includes laboratory-measured soil parameter values for the study area and three other regions of Dakshina Kannada (District), including Marpadi, Mangalore with longitude: 75.720343, latitude: 14.360019, Beluvai, Mangalore with longitude: 74.900180, latitude: 15.100230, and Attur, Mangalore with longitude: 74.820230, latitude: 15.100230 as shown in Table 1. Equation (5) is used to compute the observed variation.

$$Observed\ variations = |Derived\ values\ using\ Sentinel-2\ data - Soil\ health\ data| \tag{5}$$

Soil nutrients are more soluble in acidic soils than in neutral or slightly alkaline soils. Therefore, pH values influence nutrient availability and are considered indicators of other soil parameters (Tharavathy, 2016). To label the dataset manually, we use the soil parameter level of B, Fe, K, phosphorous (P), Mn, sulphur (S), Cu and Zn as LOW or MEDIUM or HIGH by using the value of soil pH as given in Table 2. The soil parameter level of EC, pH, OC, and N was estimated using the measured values of soil parameters as shown in Table 3.

Using manual labelling, 293 instances in the dataset are labelled as LOW, 25 as MEDIUM, and 11 as HIGH soil fertility. Table 4 shows the classification results. The RF classifier and DT classifier with a ten-fold cross-validation test achieved the highest accuracy of 98.48%, precision, recall, and F-measure of 0.985.

5 Clustering methods in soil fertility estimation

The Sentinel-2 dataset's data points for the study region were grouped in this work using a variety of clustering approaches. The dataset was labelled using clustering techniques with a predetermined number of clusters (i.e., 3). The classification accuracy of the clustered dataset was assessed using four different machine learning-based classifiers.

Figure 4 Sentinel-2 spectral bands of a selected area of study, (a) B3 (b) B4 (c) B5 (d) B8 (e) B9 (f) B11 (g) B12 (h) QA60

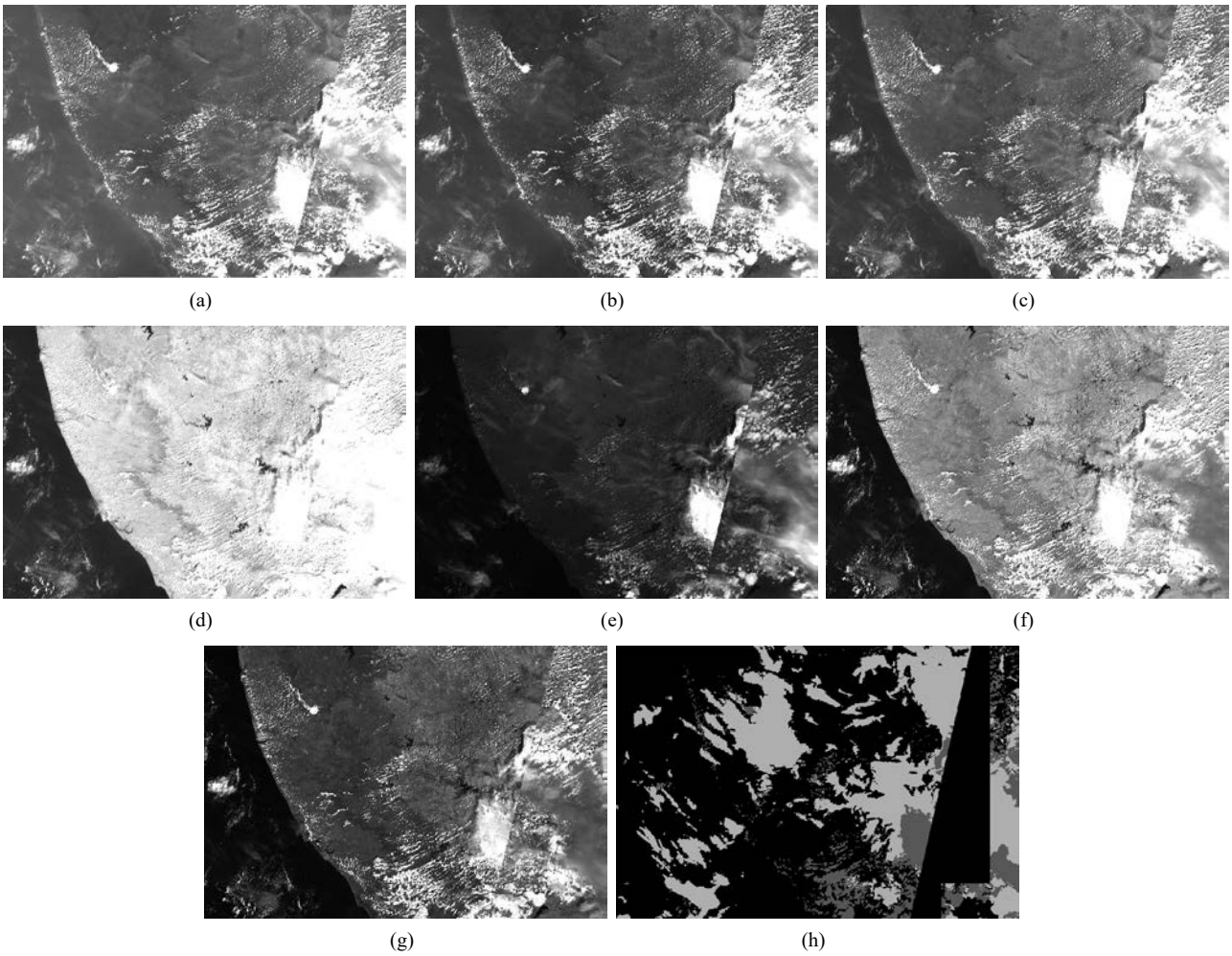


Figure 5 Combination of Sentinel-2 spectral bands used, (a) B4, B8 (b) B12, B8 (c) B3, B8 (d) B4, B12 (e) B9, B12 (f) B5, B11 (g) B12, B11 (h) B4, B3 (i) B8, B5 (see online version for colours)

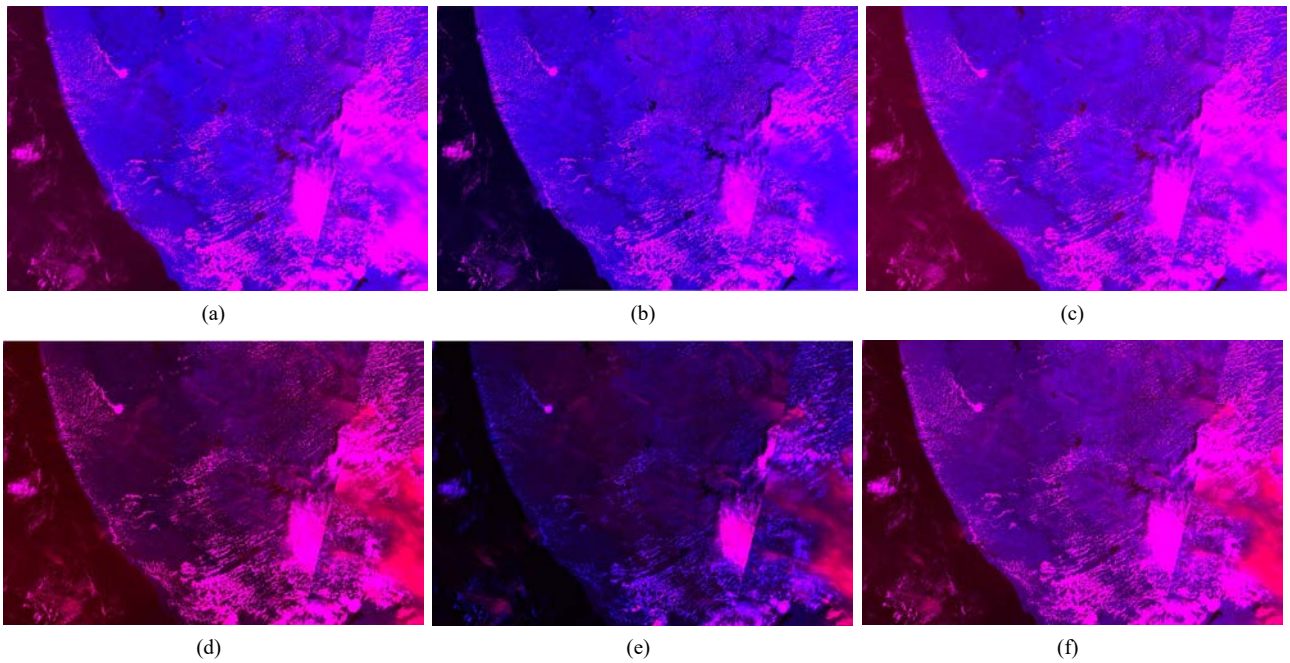


Figure 5 Combination of Sentinel-2 spectral bands used, (a) B4, B8 (b) B12, B8 (c) B3, B8 (d) B4, B12 (e) B9, B12 (f) B5, B11 (g) B12, B11 (h) B4, B3 (i) B8, B5 (continued) (see online version for colours)

(g) (h) (i)

Table 1 Comparision of derived values using Sentinel-2 with soil-health data for the study area Konaje, Marpadi, Beluvai and Attur of Mangalore Region, Karnataka (State), India

Area used	*Derived values using Sentinel-2 data*			*Soil health data*			*Observed variations (difference)*		
	EC	*pH*	*OC*	*EC*	*pH*	*OC*	*EC*	*pH*	*OC*
Konaje	0.533	6.34	2.406	0.385	6.33	2.2	0.148	*0.01*	0.206
	0.975	6.44	1.522	0.962	6.4	1.16	0.013	0.04	0.362
	1.327	6.59	0.82	1.023	6.51	0.86	*0.304*	0.08	*0.04*
	0.282	6.57	1.128	0.264	6.71	1.41	0.018	0.14	0.282
	1.979	5.36	0.956	1.91	5.87	1.071	0.069	*0.51*	0.115
	1.633	5.64	1.082	1.659	5.76	1.552	0.026	0.12	*0.47*
	1.171	6.43	0.987	1.116	6.41	0.68	0.055	0.02	0.307
	0.282	6.57	1.128	0.275	6.22	1.54	*0.007*	0.35	0.412
Marpadi	1.357	5.63	1.209	1.258	5.69	1.852	*0.099*	0.06	0.643
	1.297	6.18	0.947	1.82	6.13	1.606	0.523	0.05	*0.659*
	1.161	5.74	1.039	1.329	5.77	1.21	0.168	0.03	*0.171*
	1.153	5.85	0.822	1.347	5.85	1.397	0.194	*0*	0.575
	1.129	5.99	0.905	1.331	5.9	1.312	0.202	0.09	0.407
	0.935	8.66	1.595	1.639	8.51	1.921	*0.704*	*0.15*	0.326
Beluvai	1.615	6.53	1.146	1.118	6.79	1.9	0.497	0.26	0.754
	1.615	2.17	1.175	2.202	2.51	1.37	0.587	*0.34*	0.195
	1.615	7.89	1.066	1.263	7.63	1.98	0.352	0.26	*0.914*
	1.615	5.72	2.679	1.024	5.81	2.469	*0.591*	0.09	0.21
	1.615	6.22	1.816	1.027	6	1.7	0.588	0.22	*0.116*
	1.615	5.99	1.825	1.367	5.93	1.496	*0.248*	*0.06*	0.329
Attur	1.846	5.58	1.03	1.24	5.47	1.924	0.606	0.11	0.894
	2.089	4.36	1.133	2.12	4.84	2.239	*0.031*	*0.48*	*1.106*
	1.035	5.22	1.598	1.997	5.12	2.241	*0.962*	*0.1*	0.643
	0.944	2.64	2.483	1.817	2.47	2.29	0.873	0.17	*0.193*

Table 2 Estimation soil parameter levels based on pH value

Soil parameter	*LOW*	*MEDIUM*	*HIGH*
B	pH < 5.1 or 8 < pH < 8.5	7.5 < pH <= 8 or 8.5 < pH <= 8.75	5.1 <= pH <= 7.5 or pH >= 8.5
Fe	pH > 8	7.5 < pH <= 8.0	pH <= 7.5
K	pH < 5.5	5.5 <= pH < 5.9	pH >= 5.9
Mn	pH < 5	5 <= pH < 5.4 or 7.5 < pH <= 8	5.4 <= pH <= 7.5
P	pH < 5.5 or 8.5 < pH < 9.0	5.5 <= pH < 5.9 or 7.5 < pH <= 8.5	5.9 <= pH <= 7.5 or pH >= 9
S	pH < 5.5	5.5 <= pH < 5.9	pH >= 5.9
Cu, Zn	pH < 4.5 or pH > 8	4.5 <= pH < 5 or 7.5 < pH <= 8	5 <= pH <= 7.5

Table 3 Estimation soil parameter levels of EC, pH, OC, and N

Soil parameter	*LOW*	*MEDIUM*	*HIGH*
EC	>2.5	>1.6, <=2.5	<=1.6
pH	>8.5, <6.5	-	<=8.5, >=6.5
OC	<0.5	>=0.5, <0.75	>=0.75
N	<280	>=280, <560	>=560

By calculating the approximate distances between data point pairs and a distance threshold (T1, T2) with T1 > T2, Canopy clustering was implemented. The algorithm starts with a set of data points, eliminates each one at a time, and then repeats the process over the remaining points to produce a Canopy that includes the removed points. If the farthest points are closer to the starting point than T1, they are included in the cluster. Additionally, the point was eliminated from the set if the distance was

less than T2. The algorithm iterates until the initial set is empty, building a set of canopies, each with one or more points. By using T1 = 1.25 and T2 = 0.75, we obtained the highest clustering accuracy. The Canopy clustering resulted in 279 instances labelled LOW, 36 MEDIUM, and 14 HIGH. It was observed that 250 instances were clustered accurately, resulting in an accuracy of 75.99%. The classification results obtained after Canopy clustering are shown in Table 5. The RF classifier with a 75% split of the dataset as training data and a 25% split of the dataset as test data achieved the highest accuracy of 98.78%, precision of 0.989, recall of 0.988, and F-measure of 0.987.

Table 4 Classification using manually labelled dataset

Method used	*Classifier name*	*Accuracy*	*Precision*	*Recall*	*F-measure*
10-fold	NB	93.92%	0.940	0.939	0.939
cross-	SVM	89.06%	0.877	0.891	0.866
validation	DT	98.48%	0.985	0.985	0.985
	RF	98.48%	0.985	0.985	0.985
75% split	NB	93.42%	0.983	0934	0.953
training data:	SVM	90.24%	0.912	0.902	0.877
25% split	DT	97.56%	0.984	0.976	0.976
test data	RF	97.56%	0.984	0.976	0.976

Table 5 Data classification after Canopy clustering

Method used	*Classifier name*	*Accuracy*	*Precision*	*Recall*	*F-measure*
10-fold	NB	95.44%	0.960	0.954	0.956
cross-	SVM	90.27%	0.913	0.903	0.882
validation	DT	95.14%	0.955	0.951	0.953
	RF	97.87%	0.980	0.979	0.979
75% split	NB	95.12%	0.953	0951	0.952
training data:	SVM	91.46%	0.922	0.915	0.899
25% split	DT	96.34%	0.963	0.963	0.963
test data	RF	98.78%	0.989	0.988	0.987

Table 6 Data classification after density-based clustering

Method used	*Classifier name*	*Accuracy*	*Precision*	*Recall*	*F-measure*
10-fold	NB	96.35%	0.964	0.964	0.964
cross-	SVM	86.63%	0.881	0.866	0.860
validation	DT	97.87%	0.979	0.979	0.979
	RF	96.96%	0.970	0.970	0.970
75% split	NB	96.34%	0.964	0.963	0.964
training data:	SVM	85.37%	0.857	0.854	0.850
25% split	DT	96.34%	0.964	0.963	0.963
test data	RF	96.34%	0.964	0.963	0.963

Density-based clustering uses the local density of each data point to calculate an outlier score. If the local density of a particular data point is low compared to its neighbours, the data point is likely an outlier (Naghavi-Nozad et al., 2021). The data point with the highest local density is chosen as the cluster centre (Gu et al., 2020). Density-based clustering identified 135 instances as LOW, 148 as MEDIUM, and 46 as HIGH. It was found that 164 instances were clustered accurately, with an accuracy of 50%. The results of classification obtained after density-based clustering are shown in Table 6. RF classifier with a ten-fold cross-validation test achieved the highest accuracy of 96.96% and precision, recall, and F-measure of 0.970.

Table 7 Data classification after expectation-maximisation clustering

Method used	*Classifier name*	*Accuracy*	*Precision*	*Recall*	*F-measure*
10-fold	NB	95.44%	0.960	0.954	0.956
cross-	SVM	82.98%	0.842	0.830	0.823
validation	DT	96.96%	0.970	0.970	0.970
	RF	97.57%	0.976	0.976	0.976
75% split	NB	96.34%	0.965	0.963	0.964
training data:	SVM	74.39%	0.784	0.744	0.738
25% split	DT	95.12%	0.956	0.951	0.951
test data	RF	96.34%	0.966	0.963	0.963

Table 8 Data classification after farthest-first clustering

Method used	*Classifier name*	*Accuracy*	*Precision*	*Recall*	*F-measure*
10-fold	NB	89.67%	0.924	0.897	0.901
cross-	SVM	84.50%	0.809	0.845	0.826
validation	DT	96.96%	0.970	0.970	0.970
	RF	97.87%	0.979	0.979	0.979
75% split	NB	91.46%	0.954	0.915	0.926
training data:	SVM	74.39%	0.750	0.744	0.718
25% split	DT	91.46%	0.925	0.915	0.912
test data	RF	96.34%	0.964	0.963	0.963

Table 9 Data classification after k-means clustering

Method used	*Classifier name*	*Accuracy*	*Precision*	*Recall*	*F-measure*
10-fold	NB	95.74%	0.958	0.957	0.958
cross-	SVM	84.80%	0.866	0.848	0.839
validation	DT	95.14%	0.951	0.951	0.951
	RF	96.05%	0.960	0.960	0.960
75% split	NB	95.12%	0.952	0.951	0.951
training data:	SVM	82.93%	0.839	0.829	0.825
25% split	DT	95.12%	0.952	0.951	0.951
test data	RF	95.12%	0.952	0.951	0.951

Table 10 Data classification after FCM clustering

Method used	*Classifier name*	*Accuracy*	*Precision*	*Recall*	*F-measure*
10-fold	NB	81.76%	0.947	0.818	0.813
cross-	SVM	86.93%	0.875	0.869	0.852
validation	DT	98.18%	0.982	0.982	0.982
	RF	97.57%	0.976	0.976	0.975
75% split	NB	80.49%	0.834	0.805	0.800
training data:	SVM	84.15%	0.869	0.841	0.808
25% split	DT	98.78%	0.988	0.988	0.988
test data	RF	97.56%	0.976	0.976	0.976

Expectation-maximisation clustering uses the probability that each data point is present in either cluster. This approach resulted in 126 instances being classified as LOW, 161 as MEDIUM, and 42 as HIGH. This method of clustering correctly clustered 155 instances, with an accuracy of 47%. Table 7 displays the classification results that were achieved using expectation-maximisation clustering. The RF classifier attained the best accuracy of

97.57% and precision, recall, and F-measure of 0.976 using a ten-fold cross-validation test.

Table 11 Data classification after proposed Canopy centre-based FCMs clustering

Method used	*Classifier name*	*Accuracy*	*Precision*	*Recall*	*F-measure*
10-fold	NB	97.26%	0.973	0.973	0.973
cross-	SVM	95.14%	0.956	0.951	0.951
validation	DT	99.39%	0.994	0.994	0.994
	RF	99.69%	0.997	0.997	0.997
75% split	NB	96.34%	0.965	0.963	0.964
training data:	SVM	92.68%	0.937	0.927	0.927
25% split	DT	98.78%	0.988	0.988	0.988
test data	RF	98.78%	0.988	0.988	0.988

Farthest-first clustering selects a random data point as the initial cluster centre. During the cluster assignment phase, the data point farthest from the first centre is chosen as the new centre. This process is repeated until the 'k' number of centroids has been chosen. Each remaining data point is assigned to the cluster characterised by the centroid nearest to the data point, and the algorithm terminates. Farthest-first requires a single pass to cluster a set of data points (Devi et al., 2020). This clustering method resulted in 216 instances of LOW, 96 of MEDIUM, and 17 of HIGH soil fertility. This method of clustering clustered 182 instances accurately, with an accuracy of 55.32%. The results of classification obtained after farthest-first clustering are shown in Table 8. The RF classifier with a ten-fold cross-validation test of the dataset obtained the highest accuracy of 97.87%, precision, recall, and F-measure of 0.979.

The k-means clustering algorithm selects k random data points from the cluster centroids, where k is equal to the number of clusters, and uses a distance metric (usually Euclidean) to assign all the points with the closest distances to the centroid. The algorithm iteratively computes the centroids of newly formed clusters and assigns the remaining data points to the cluster with the closest centroid until clusters are stable (Wang and Kumar, 2019). It is difficult to achieve ideal clusters since the k-means clustering is sensitive to outliers (Guo et al., 2021). Using k-means clustering, 137 instances were classified as having low soil fertility, 146 as having medium soil fertility, and 46 as having high soil fertility. This method of clustering has a 50.46% and correctly grouped 166 instances. Table 9 displays the classification results that were reached using k-means clustering. RF classifier with a ten-fold cross-validation test performed better with an accuracy of 96.05% and with precision, recall, and F-measure of 0.960.

FCMs clustering is an unsupervised machine learning algorithm that assigns data points to clusters, with points belonging to the same cluster being as similar as possible, and each data point may belong to more than one cluster (Chen et al., 2022). The FCM clustering can improve the classification speed (Chen et al., 2022), whereas it is sensitive to initial cluster centroids (Xue et al., 2016). The clustering is achieved based on the minimisation of the objective function given in equation (6).

$$\sum_{i=1}^{N}\sum_{j=1}^{k}\mu_{ij}^{m}\left|x_i - c_j\right|^2 \tag{6}$$

where N is the number of objects and k is the number of clusters, μ_{ij} is the degree of membership of instance x_i in the j^{th} cluster, m is the fuzzy parameter which indicates the degree of fuzzy overlap, x_i indicates i^{th} instance, and c_j represents the centre of the j^{th} cluster. Initially, μ_{ij} is set randomly, then the c_j and updated μ_{ij} are calculated by using equations (7) and (8), respectively.

$$c_j = \frac{\sum_{i=1}^{N}\mu_{ij}^{j}x_i}{\sum_{i=1}^{N}\mu_{ij}^{j}} \tag{7}$$

$$\mu_{ij}^{m} = \frac{1}{\sum_{k=1}^{N}\left(\frac{|x_i - c_j|}{|x_i - c_k|}\right)^{\frac{1}{m-1}}} \tag{8}$$

This clustering technique resulted in the classification of 155 instances as LOW, 90 as MEDIUM, and 84 as HIGH. The highest accuracy attained by this clustering, with the correct clustering of 145 cases, was 44%. In Table 10, the classification results of FCM clustering are depicted. The DT classifier achieved the greatest accuracy of 98.78%, precision, recall, and F-measure of 0.988 using a 75% split as training data and a 25% split of the dataset as test data.

6 Proposed Canopy centre-based FCMs clustering method

The state-of-the-art FCM clustering uses random cluster centres, which will limit the accuracy of clustering. Hence, the proposed approach selects the cluster centres using the best-performing clustering approach. From the experimental results, it was observed that Canopy clustering achieved better accuracy. Hence, to improve the clustering accuracy, we proposed a Canopy centre-based FCMs clustering method as depicted in Algorithm 1.

Canopy centres were computed as initial centroids for the FCMs clustering algorithm with the fixed number of clusters and fixed threshold values T1 = 1.25 and T2 = 0.75. Soil fertility can be LOW, MEDIUM, or HIGH. Hence the number of clusters was fixed to 3 (k = 3). We selected a value of fuzzy parameter m, such that 1.1 < m < 2, and the algorithm obtained better accuracy for m = 1.7. The membership matrix, M (i.e., membership function), was initialised using random values and updated by calculating the Euclidean distance between a pair of data points and cluster centres using equation (9).

$$M_{ij} = \frac{1}{\sum_{k=1}^{N}\left(\frac{\langle x_i - c_j\rangle^2}{\langle x_i - c_k\rangle^2}\right)^{\frac{1}{m-1}}} \tag{9}$$

where M_{ij} indicates degree to which an observation x_i belongs to cluster c_j. The value M_{ij} inversely proportional to the distance from x to cluster centre.

Algorithm 1 Canopy centre-based FCMs clustering

Initialise X with the given dataset and assign the number of clusters, k
Assign fuzzy parameter, m with 1 < m < 2
Initialise membership matrix M with random values [0, 1]

procedure CANOPYFUZZYCMEANSCLUSTERING(X, k, m, M)
  $cur\_iter \leftarrow 0$
  $cluster\_labels \leftarrow \{\}$
  $cluster\_centres \leftarrow$ CALCULATECANOPYCENTRE(X)
  while cur_iter < MAX_ITER **do**
    $M \leftarrow$ UPDATEMEMBERSHIPMATRIX($M, cluster\_centres$)
    $clustercentres \leftarrow$ CALCULATECANOPYCENTRE(M)
    for i = 1 to N **do**
      $idx \leftarrow Indexof\_minimum\_M[i]$
      $cluster\_labels = cluster\_labels \cup idx$
    $cur\_iter \leftarrow cur\_iter + 1$
  return $cluster\_labels$
procedure CALCULATECANOPYCENTRE(X)
  Initialise threshold T1, T2 such that T1 > T2
  $canopies = \{\}$
  $dist[i, j] \leftarrow$ EUCLIDEANDISTANCE$(x_i, x_j), \forall (x_i, x_j) \in X$
  $Canopy\_points = (x_i, x_j)$
  while $Canopy\_points \neq \{\}$ **do**
    $point \leftarrow pop(Canopy\_points)$
    $i \leftarrow Length(canopies)$
  if $dist[point] < T1$ **then**
    $canopies[i] \leftarrow point$
  if $dist[point] < T2$ **then**
    $X = X - point$
  return $canopies$
procedure UPDATEMEMBERSHIPMATRIX(M, cluster_centres)
  $p \leftarrow \frac{2}{m-1}$
  for i = 1 to N **do**
    for j = 1 to k **do**
      $distances = X[i] - cluster\_centres[j]$
    for j in k **do**
      for q = 1 to k **do**
        $sum = sum + \frac{distances[j]^p}{distances[q]^p}$
        $M[i] \leftarrow \frac{1}{sum}$
  return M

The algorithm used a maximum of 100 iterations to obtain optimal clustering. Using FCM clustering, the fuzziness of the datapoint belonging to more than one cluster is avoided by selecting a maximum value from the membership function. But, soil fertility depends on the level of each soil parameter. The low fertility level of any soil chemical parameters makes the soil less fertile. Thus, instead of selecting maximum membership value, the algorithm selects minimum membership value. On overlap, the datapoint falls to a cluster with a minimum cluster number.

The proposed clustering technique produced 254 instances of LOW, 60 instances of MEDIUM, and 15 instances of HIGH fertile soil. With an accuracy of 78.42%, this approach correctly clustered 258 instances. The comparison of the accuracy of all clustering techniques used in the study is depicted in Figure 6.

The proposed method achieved the highest clustering accuracy as compared to other clustering methods. The results of applying classification are presented in Table 11. The RF classifier with ten-fold cross-validation of the dataset obtained the highest accuracy of 99.69%, precision, recall, and F-measure of 0.977.

Figure 6 Comparison of accuracy of clustering techniques

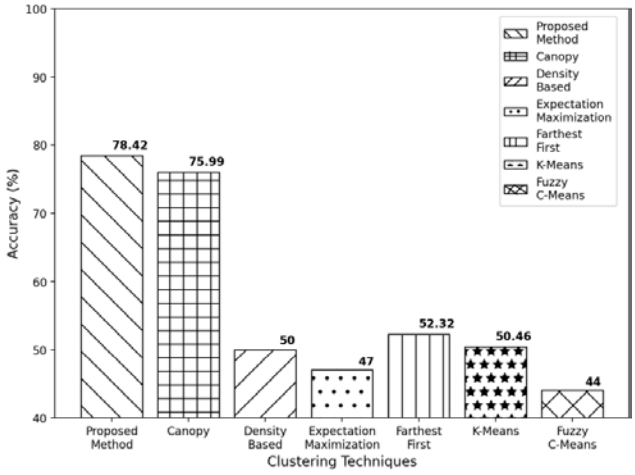


Figure 7 Comparision of clustering techniques based on NB classification

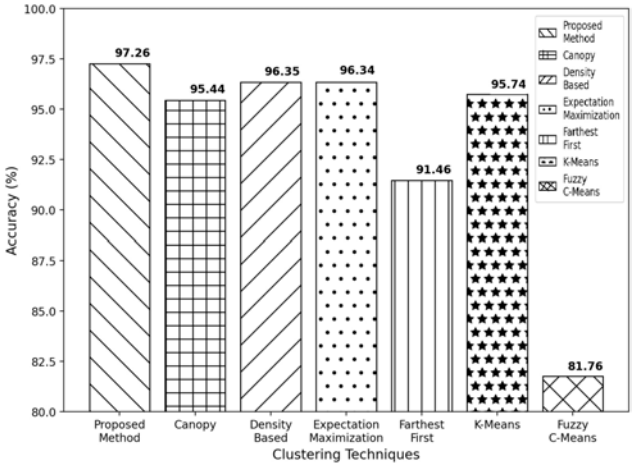
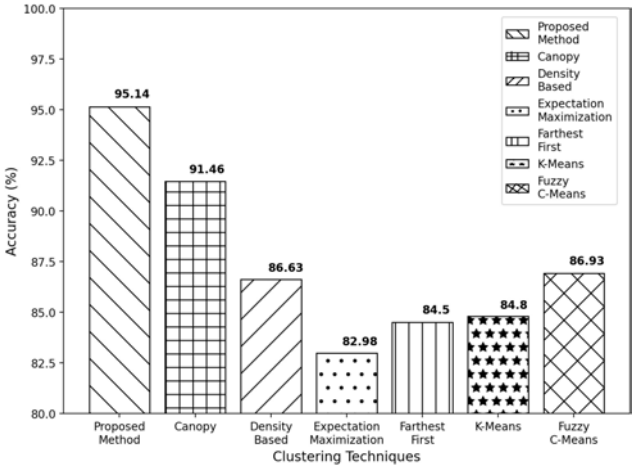


Figure 8 Comparision of clustering techniques based on SVM classification



Figures 7, 8, 9 and 10 indicate the accuracy of different clustering techniques using NB, SVM, DT, and RF, respectively.

Figure 9 Comparision of clustering techniques based on DT classification

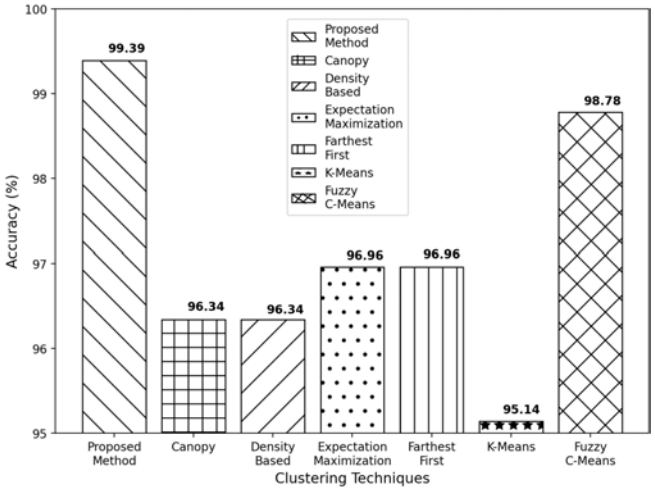
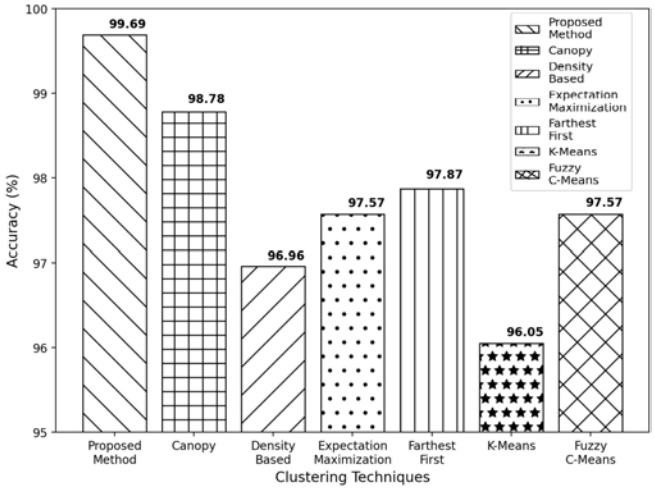


Figure 10 Comparision of clustering techniques based on RF classification



7 Conclusions

Estimating the site-specific soil fertility is essential to determine the most cost-effective fertiliser application. The accurate prediction of fertilisers aids in lowering environmental pollution caused by excessive fertilisation. It is not economical to predict soil fertility using laboratory measurements. Furthermore, laboratory measurements leave behind chemical residues and take extra time. This study effort employed remotely sensed Sentinel-2 spectral bands to predict soil parameters such as *EC*, *pH*, *OC* and *N* in order to overcome these limitations. The data points were clustered using a variety of cutting-edge clustering techniques. It was found that Canopy clustering achieved a clustering accuracy of 75.99%, and by utilising an RF classifier with ten-fold cross-validation, Canopy clustering obtained a classification accuracy of 98.78%. It was observed that using the proposed Canopy centre-based FCMs clustering achieved the highest clustering accuracy of 78.42%, and by utilising an RF classifier with a ten-fold cross-validation proposed approach obtained a classification accuracy of 99.69%. Decisions on soil fertility are more precise using the proposed clustering technique. Farmers are, therefore, able to know the level of fertility of their soil at any given time and apply fertiliser in line with the crop being grown on their farm. This enables the farmers to increase their agricultural yields and profits. Increased agricultural production boosts export or business, thus helping the agriculture industry. Future research might use real-time soil data collected during crop development to increase accuracy. Sentinel-2 gathers the data every five days, and frequent revisits are necessary to increase the accuracy. Proximate soil sensors allow for the dynamic collection of soil data to determine the variation in soil fertility based on crop development and environmental factors.

References

Aksoy, S., Yildirim, A., Gorji, T., Hamzehpour, N., Tanik, A. and Sertel, E. (2022) 'Assessing the performance of machine learning algorithms for soil salinity mapping in Google Earth Engine platform using Sentinel-2A and Landsat-8 OLI data', *Advances in Space Research*, Vol. 69, No. 2, pp.1072–1086.

Al-Gaadi, K.A., Tola, E., Madugundu, R. and Fulleros, R.B. (2021) 'Sentinel-2 images for effective mapping of soil salinity in agricultural fields', *Current Science*, Vol. 121, No. 3, p.384.

Chen, G., Li, X., Gong, W. and Xu, H. (2020) 'Recognition of the landslide disasters with extreme learning machine', *International Journal of Computational Science and Engineering*, Vol. 21, No. 1, pp.84–94.

Chen, H., Das, S., Morgan, J.M. and Maharatna, K. (2022) 'Prediction and classification of ventricular arrhythmia based on phase-space reconstruction and fuzzy C-means clustering', *Computers in Biology and Medicine*, Vol. 142, p.105180.

Chougule, A., Jha, V.K. and Mukhopadhyay, D. (2019) 'Crop suitability and fertilizers recommendation using data mining techniques', in *Progress in Advanced Computing and Intelligent Engineering*, pp.205–213, Springer, Singapore.

Dasgupta, S., Chakraborty, S., Weindorf, D.C., Li, B., Silva, S.H.G. and Bhattacharyya, K. (2022) 'Influence of auxiliary soil variables to improve PXRF-based soil fertility evaluation in India', *Geoderma Regional*, Vol. 30, p.e00557.

Devi, R.D.H., Bai, A. and Nagarajan, N. (2020) 'A novel hybrid approach for diagnosing diabetes mellitus using farthest-first and support vector machine algorithms', *Obesity Medicine*, Vol. 17, p.100152.

Divakar, M.S., Elayidom, M.S. and Rajesh, R. (2022) 'Design and implementation of an efficient and cost effective deep feature learning model for rice yield mapping', *International Journal of Computational Science and Engineering*, Vol. 25, No. 2, pp.128–139.

Du, H., Hao, Y. and Wang, Z. (2022) 'An improved density peaks clustering algorithm by automatic determination of cluster centres', *Connection Science*, Vol. 34, No. 1, pp.857–873.

FAO (2022) *Soil Fertility* [online] https://www.fao.org/global-soil-partnership/areas-of-work/soil-fertility/en/ (accessed 30 August 2022).

Gholizadeh, A., Žižala, D., Saberioon, M. and Borůvka, L. (2018) 'Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging', *Remote Sensing of Environment*, Vol. 218, pp.89–103.

Gorji, T., Yildirim, A., Hamzehpour, N., Tanik, A. and Sertel, E. (2020) 'Soil salinity analysis of Urmia Lake Basin using Landsat-8 OLI and Sentinel-2A based spectral indices and electrical conductivity measurements', *Ecological Indicators*, Vol. 112, p.106173.

Gu, X., Peng, J., Cheng, Y., Zhang, X. and Liu, K. (2020) 'Energy replenishment optimisation via density-based clustering', *International Journal of Computational Science and Engineering*, Vol. 21, No. 2, pp.271–280.

Guo, Y., Wu, Y., Zhang, X., Bo, A. and Li, X. (2021) 'The FRCK clustering algorithm for determining cluster number and removing outliers automatically', *International Journal of Computational Science and Engineering*, Vol. 24, No. 5, pp.485–494.

Hengl, T., Miller, M.A., Križan, J., Shepherd, K.D., Sila, A., Kilibarda, M., Antonijević, O., Glušica, L., Dobermann, A., Haefele, S.M. and McGrath, S.P. (2021) 'African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning', *Scientific Reports*, Vol. 11, No. 1, pp.1–18.

Khanal, S., Fulton, J., Klopfenstein, A., Douridas, N. and Shearer, S. (2018) 'Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield', *Computers and Electronics in Agriculture*, Vol. 153, pp.213–225.

Kim, H., Kim, H.K. and Cho, S. (2020) 'Improving spherical k-means for document clustering: fast initialization, sparse centroid projection, and efficient cluster labeling', *Expert Systems with Applications*, Vol. 150, p.113288.

Li, X., Wang, X. and Wu, L. (2017) 'System architecture of coastal remote sensing data mining and services based on cloud computing', *International Journal of Computational Science and Engineering*, Vol. 15, Nos. 3–4, pp.267–276.

Ma, Z. and Li, B. (2020) 'A DDoS attack detection method based on SVM and K-nearest neighbour in SDN environment', *International Journal of Computational Science and Engineering*, Vol. 23, No. 3, pp.224–234.

Mashaba-Munghemezulu, Z., Chirima, G.J. and Munghemezulu, C. (2021) 'Modeling the spatial distribution of soil nitrogen content at smallholder maize farms using machine learning regression and Sentinel-2 data', *Sustainability*, Vol. 13, No. 21, p.11591.

Naghavi-Nozad, S.A., Haeri, M.A. and Folino, G. (2021) 'SDCOR: scalable density-based clustering for local outlier detection in massive-scale datasets', *Knowledge-Based Systems*, Vol. 228, p.107256.

NRCS-USDA (2020) *Soil Electrical Conductivity* [online] https://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcs142p2_053280.pdf (accessed 25 January 2022).

Priya, R. and Ramesh, D. (2020) 'ML based sustainable precision agriculture: a future generation perspective', *Sustainable Computing: Informatics and Systems*, Vol. 28, p.100439.

Samandi, V. and Mukhopadhyay, D. (2021) 'Workflow scheduling in cloud computing environment with classification ordinal optimisation using SVM', *International Journal of Computational Science and Engineering*, Vol. 24, No. 6, pp.563–571.

Sentinel-2 (2020) *Sentinel-2 MSI: Multispectral Instrument, Level-1C* [online] https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2, (accessed 23 October 2021).

Shahmoradi, Z. and Lee, T. (2022) 'Optimality-based clustering: an inverse optimization approach', *Operations Research Letters*, Vol. 50, No. 2, pp.205–221.

Sirsat, M.S., Cernadas, E., Fernández-Delgado, M. and Khan, R. (2017) 'Classification of agricultural soil parameters in India', *Computers and Electronics in Agriculture*, Vol. 135, pp.269–279.

Sirsat, M.S., Cernadas, E., Fernández-Delgado, M. and Barro, S. (2018) 'Automatic prediction of village-wise soil fertility for several nutrients in India using a wide range of regression methods', *Computers and Electronics in Agriculture*, Vol. 154, pp.120–133.

Soil-Health Data (2021) *Soil Health Card India, Nutrient Status-Sample Wise (For Geo Coordinates Updation)* [online] https://soilhealth.dac.gov.in/PublicReports/nutrientstatussamplesurveywise (accessed 1 July 2021).

Tayeb, M.S. and Fizazi, H. (2020) 'A new neural architecture for feature extraction of remote sensing data', *International Journal of Computational Science and Engineering*, Vol. 21, No. 1, p.95.

Tharavathy, N.C. (2016) 'A study on soil characteristics in urban and rural areas of Mangalore, Karnataka', *International Journal of Research in Environmental Science*, Vol. 2, No. 2, pp.5–8.

Vaudour, E., Gomez, C., Fouad, Y. and Lagacherie, P. (2019) 'Sentinel-2 image capacities to predict common topsoil properties of temperate and Mediterranean agroecosystems', *Remote Sensing of Environment*, Vol. 223, pp.21–33.

Wang, K. and Kumar, P. (2019) 'Characterizing relative degrees of clumping structure in vegetation Canopy using waveform LiDAR', *Remote Sensing of Environment*, Vol. 232, p.111281.

Wang, B., Waters, C., Orgill, S., Cowie, A., Clark, A., Liu, D.L., Simpson, M., McGowen, I. and Sides, T. (2018) 'Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of Eastern Australia', *Ecological Indicators*, Vol. 88, pp.425–438.

Wang, C., Zhao, T. and Mo, X. (2019) 'The extraction of security situation in heterogeneous log based on STR-FSFDP density peak cluster', *International Journal of Computational Science and Engineering*, Vol. 20, No. 3, pp.387–396.

WEKA (2021) *WEKA: Machine Learning Software in Java* [online] https://waikato.github.io/weka-wiki/downloading_weka/ (accessed 14 November 2021).

Wu, Y., Peng, X., Mohammad, N. and Yang, H. (2021) 'Research on fuzzy clustering method for working status of mineral flotation process', *International Journal of Embedded Systems*, Vol. 14, No. 2, pp.133–142.

Xu, Y., Smith, S.E., Grunwald, S., Abd-Elrahman, A., Wani, S.P. and Nair, V.D. (2018) 'Estimating soil total nitrogen in smallholder farm settings using remote sensing spectral indices and regression kriging', *Catena*, Vol. 163, pp.111–122.

Xue, Y., Zhao, B. and Ma, T. (2016) 'Performance analysis for clustering algorithms', *International Journal of Computing Science and Mathematics*, Vol. 7, No. 5, p.485.

Yang, R.M. and Guo, W.W. (2019) 'Using time-series Sentinel-1 data for soil prediction on invaded coastal wetlands', *Environmental Monitoring and Assessment*, Vol. 191, No. 7, pp.1–14.

Ye, Y., Sun, X., Liu, M., Zhao, Z., Zhang, X. and Wu, H. (2018) 'The remote farmland environment monitoring system based on ZigBee sensor network', *International Journal of Computational Science and Engineering*, Vol. 17, No. 1, pp.25–33.

Zhang, G., Zhang, C. and Zhang, H. (2018) 'Improved K-means algorithm based on density Canopy', *Knowledge-based Systems*, Vol. 145, pp.289–297.

Zhu, G., Li, X., Zhang, S., Xu, X. and Zhang, B. (2022) 'An improved method for k-means clustering based on internal validity indexes and inter-cluster variance', *International Journal of Computational Science and Engineering*, Vol. 25, No. 3, pp.253–261.