Int. J. Nonlinear Anal. Appl.
Volume 11, Special Issue, Winter and Spring 2020, 527-541
ISSN: 2008-6822 (electronic)
http://dx..doi.org/10.22075/ijnaa.2020.47500

Terrain Mapping of LandSat8 Images using MNF and Classifying Soil Properties using Ensemble Modelling

K. Lavanya[a], Ahmed J Obaid[b], I. Sumaiya Thaseen[c], Kumar Abhishek[d], Khushboo Saboo[a], Rucha Paturkar[a]

[a]*School of Computing Science and Engineering, Vellore Institute of Technology, Vellore.*
[b]*Department of Computer Science, Faculty of Computer Science and Mathematics, University of Kufa, Iraq.*
[c]*School of Information Technology and Engineering, Vellore Institute of Technology Vellore*
[d]*Department of Computer Science, National Institute of Technology, Patna, Patna-800005.*

Abstract

Traditional technique for determining the soil texture and other soil properties is performed in laboratory which is a time consuming task. In this paper, machine learning algorithms are deployed to classify the soil texture and its properties without any intervention of laboratory equipment using the satellite images recorded by Landsat 8. These images are used to extract the terrain properties of the region which is integrated with weather data for the specific region and the vegetation index which are the major factors affecting the soil condition. A major aim of this paper is to design a robust technique for extracting, transforming Landsat images to numerical data and pre-processing the data for classifying the soil property. Minimum Noise Fraction (MNF) is utilized to segregate and remove noise from the Landsat images for subsequent processing. A significant amount of noise is present in the raw data which affects the accuracy of the analysis. Terrain features are extracted after noise removal from the MNF transformed images and merged with the weather data, and vegetation index for a period of time and then classified using voting classifier of the ensemble modeling or analysis of the soil texture of the region. The voting is performed by integrating the results of logistic regression, support vector machine and decision tree. With this study, the consolidated dependence of the soil texture on the environmental factors is analyzed and a cross validation accuracy of 94.44% is obtained.

Keywords: Decision Tree, Digital Soil Mapping, Ensemble Modelling, Landsat8, MNF, Noise Removal, Soil Texture, SVM, Terrain Analysis and Voting Classifier.

*K. Lavanya
Email addresses: ahmedj.aljanaby@uokufa.edu.iq (Ahmed J Obaid), kumar.abhishek@nitp.ac.in (Kumar Abhishek)

Received: February 2020 *Revised:* October 2020

1. Introduction

Accurate and detailed terrain and soil information is essential for efficient and sustainable land use for environment modelling. Soil Information prediction helps each sector to plan and prepare for the coming future and present. If the soil conditions are known then proper crops on the land will enrich the soil. Contrary to the traditional method of finding the composition and texture of the soil which is done in the laboratory, the proposed method takes input as the environmental factors and gives the terrain texture as the classification output.

Traditionally soil mapping was a manual process that would be based on ground surveys, soil sampling and laboratory analysis of the collected soil samples. Naturally this process was an expensive and when this was done on a nation-wide process it turned out to be a very time-consuming procedure. With change in climatic conditions the whole procedure would have to be repeated to assess the effects of rains or any thunder storm in the region has affected the soil and crop production in the region.

Digital soil mapping was introduced to overcome the above bottlenecks, by using high resolution images of the terrain as the data sources. This technique is evolving with time by increasing the terrain coverage to get information from the inaccessible areas. Remote sensing images have proved to be very fruitful data sources for digital soil mapping.

High resolution MSI images from Landsat – 8 of the terrain for the region of Vellore can assist in deciding the fertile soils and assist in maximizing the agricultural productions. The soil categories present in the Vellore region are Sandy, Sandy Loam and Clay Loam. The 4 factors taken into consideration that affect the texture of the soil are:

1. Terrain Properties like Slope, Aspect, Curvature
2. Weather Data like humidity, temperature precipitation
3. Normalized Difference Vegetation Index
4. Soil properties

Reducing noise from images is still considered a challenge in the field of image analysis and processing. The Landsat (Land Remote Sensing Satellite System) space mission was launched on July 23, 1972. It is a series of Earth observing satellites that are managed by the US Geological Survey and NASA jointly. The satellite has been providing us with images of the earth's land surface. There are 8 Landsat satellites since 1972.

Another such Space mission named Sentinal-1 was launched by the European Space Agency on 3rd April 2014 as a part of the Copernicus program. The Sentinel-1 provide us with Level-1 products such as Ground Range Detected (GRD) and Single Look Complex (SLC). The Sentinal-1 data are widely used in applications like ship detection, sea ice detection, and wind speed retrieval. The major issue with the data collection in such cases is false alarms generated due to sea beams sea-clutter and also creates data gaps. This is also referred to as GRD border noise and traditional noise removal techniques are not completely effective in removing noise over water bodies [1].

Soft computing methods have a wide range of applications [2, 3, 4].The Hyperspectral data are divided into narrow bands due to the low signal-to-noise ratio (SNR). Thus, there is a need of smoothing techniques to remove the noise and its effects on the recorded images. Minimum Noise Fraction (MNF) and Principal Component Analysis (PCA) are the two most popular algorithms for feature reduction because of their multivariate nature. MNF is a denoising technique that is a double PCA to simplify the Hyper Spectral Images (HSI) and to remove noise from hyperspectral

images. It creates a data cube of the image containing noise and transforms it to the required output channel. The images are arranged in increasing order of noise levels. The use of MNF as a denoising technique has given some promising results in the past. When MNF is used in classifying the airborne hyperspectral data on two case studies on Aosta Valley Region and Lodi (northern Italy) [5] the inter-class classification is improved and intra-class differences are highlighted. Using MNF processed data the accuracy of the inter-class classification increased from 0.36 to 0.68 as the target was noise-free.

A new MNF technique was introduced for hyperspectral data based on dimensionality reduction and thus reduced the distortion in the brightness gradient in the cross direction and improved classification mapping.

The data from all the satellite and their operational techniques are collected and projected as Landsat collection 1 and are known as Level-1 Landsat products. The sensors installed on the Landsat Satellites were designed to collect data over multiple frequencies across the electromagnetic spectrum. The Noise from the Landsat8 Images can be removed by first transforming those images to create the MNF space, smoothing the important data bands from the space and rejecting the components with maximum noise and map the space to the original image with less noise.

The features like relative elevation (Hz), Terrain wetness index (TWI), slope (β), Stream power index (SPI), Horizontal curvature (Ch) and Vertical curvature (Cv) from these transformed images are to be extracted using QGIS software. The extracted features will aid us in the further classification.

Along with the above terrain properties that were extracted from the pre-processed raster images on a monthly basis for 3 years, the weather data like precipitation, humidity, maximum and minimum temperature for the respective month. Deep Learning technique ensemble modelling is used for classification of the soil texture on the basis of above collected data.

This paper is structured as follows: Section 2 analyzes the related work. The background techniques and the study data are detailed in section 3. The proposed ensemble model is detailed in section 4. Results are determined in section 5. The paper is concluded in section 6.

2. RELATED WORK

Noise removal serves as one of the crucial steps for improving the quality of an image for visual interpretation. Effective de-noising can be achieved by filtering out unwanted signals [6].

Remote Sensing Images, such as the ones from Landsat8, are prone to a variety of types of noise. This noise in any image occurs because of incorrect process of image acquisition process which results in improper and modified pixel values and thus they do not reflect the true intensities of the real image and result in wrong input values which leads in wrong predictions. The electronic transmission also introduces noise. The results from the study conducted [6] authorized the insensitivity of random projection concerning impulse noise. Thus, in the domain of noise reduction, random projection proves a promising alternative to some existing techniques.

The PCA is determined in a manner such that the largest possible variance is assumed as the first principal component, and the next highest variances in order are considered to be the succeeding components with the restraint that it is orthogonal to every component preceding it. These components are named as the eigenvectors of the covariance matrix for the input data. In summary, the evaluation demonstrated [7] shows that the MNF-based methods are superior to PCA-based methods for signal-dependent noise and the PCA-based methods perform good in comparison to MNF-based methods for Gaussian white noise for HSI denoising.

A study was conducted [8] for a noise level on-orbit estimation utilizing early images over ground homogeneous sites for the Operational Land Imager (OLI) onboard Landsat 8. The OLI images for all the nine bands identified at different radiance levels were critical because if spectral radiance increases, SNR also increases. The dedicated study is on surface covers which include water bodies like deep oceans, vegetation areas and lakes because such land covers are disseminated consistently in the captured images over a wide area. The uncertainty of image application is determined by the noise level in radiometric standardization of remotely sensed images which is a crucial factor. In addition, aging sensors should be considered and improve the image quality by analyzing the time deviation of noise and to pixel-to-pixel radiometric standardization [9].

To avoid the laboratory analysis of the soil structure and its components which is tedious and time consuming a new systematic technique was created to detect and classify the soil data by image segmentation and multivariate image analysis (MIA). The MIA includes correlation of the digital images using PCA and Partial Least Squares (PLS) regression approach. The number of soil particles in the digital images of the soil data is estimated by MIA [7].

Principal component analysis (PCA) is another technique that can be used for hyperspectral image de-noising. PCA contains mutually uncorrelated data which results in linear groupings of the hyperspectral pixel radiance and maximum variance but it does not consider image noise and MNF increases the resultant image quality by maximizing the SNR rather than increasing the variance.

The accuracy of digital soil mapping on hyper temporal data using JM distance, K-T transformation and PCA was 85.18% and a kappa coefficient of 0.772 was achieved (Haoxuan Yang). The study was done using mono temporal classification characteristics like wetness, brightness, the greenness of the soil after the K-T transformation. The results were compared to the Hyper-Temporal classification in the same region for the same images. The images were further compressed for rich data and smaller images and then were reduced by PCA.

The study [5] evaluated the use of the MNF transform on airborne hyperspectral data for improving the classification accuracy. Two various Multispectral Infrared Visible Imaging Spectrometer (MIVIS) data sets were deployed, for inter-class and intra-class detection. The use of MNF-processed samples improves inter-class classification and also determines intra-class differences. In addition, the original data is sorted in a new structure by MNF such that the attributes indicated by low SNR are limited in a few bands. The heterogeneity of each different dataset [5] is emphasized by the structure of data distribution.

3. BACKGROUND

3.1. Study Area

This study was conducted for the city of Vellore shown in Figure 1 located in the Indian state of Tamil Nadu. It has an area of about 87.91 km2 and lies on the co-ordinates 12.9165° N, 79.1325° E. Out of the total area of 215.60 acres the cultivable soil or land is total of 175 acres. Vellore District is divided into total of 7 talukas, 753 panchayats. Major part of the Vellore region falls in the river basin of Palar river. The soil present throughout Vellore district is classified in three types Sandy Loam, Sandy and Clay Loam. Rainfall received in Vellore is due to both southwest and northwest monsoons and annual normal is 948.8 mm.

The whole district can be classified into two major physiographic divisions: Plain regions in the eastern part and Hilly terrain in southwestern and eastern parts. At the highest elevations red loamy soils can also be found whereas the valley areas are occupied by black cotton soils.

The major crop that is cultivated in the Palar Basin is paddy. Other crops like rice, maize, ragi and cotton etc., are also extensively cultivated here. The climatic conditions around the study area

is semi-arid in nature the mean daily temperature recorded for the Vellore district are 18.2 to 36.8° C and the highest temperature across the region is recorded in the months of May and June. The study area is dominated by Vermiculite, Quartz, Black Granite ores.

Figure 1: Vellore Region Map

3.2. Input data

Satellite spectral data

Multi-spectral data from Landsat8 was used in this study. A Multi spectral image captures image data for a specific region at specific wavelengths and are separated by filters. Multi-spectral imaging is known to measure light in a relatively small number of bands, typically 3 to 15. The image data is captured using the light emission by objects. The Landsat 8 satellite was launched on February 11, 2013 from Vandenberg Air Force Base, California. It orbits in a near polar orbit at an altitude of 7,05,000 m that is inclined at 98.2 degrees. One circle around the Earth is completed in 99 minutes.

The Landsat 8 satellite contains two devices namely Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS). The data for the visible bands is gathered in OLI near short wave infrared spectral bands and a panchromatic band with a precision over 12-bit dynamic range. The 12-bit data is then scaled to 16-bit integers and the Level-1 spectral images are delivered. The data for a span of 3 years was collected from the United States Geological Survey (USGS) [10] and downloaded in a GeoTIFF file format for further processing. Each file consists of 12 TIFF files, one for each band. These files contain raster data related to each of the bands.4 out of the 12 band raster images were used in the analysis. Raster images were pre-processed by removing the Noise from the images using Minimum Noise Fraction (MNF).

Climatic variables

Climatic variables such as precipitation, humidity, minimum temperature, maximum temperature and wind speed were obtained for a 3-year period. This weather data was integrated with Terrain data. This data was taken from NASA's Prediction of Worldwide Energy Resource (POWER) [11] for Vellore region. The file was pre-processed and a month wise average was calculated for each of the attributes and displayed in the Figure 2.

	A	B	C	D	E	F	G
1	YEAR	MO	Precipitation	Relative Humidity	T2M_MAX	T2M_MIN	Wind Speed Rang
2	2017	1	0.998709677	66.16516129	30.85806	17.41032	4.748387097
3	2017	2	0.003571429	54.42714286	33.92571	17.19393	6.315357143
4	2017	3	0.527	52.65533333	37.36233	22.22233	5.360333333
5	2017	4	0.195	43.126	41.204	25.33367	5.299
6	2017	5	1.670967742	53.13387097	38.36226	26.17323	5.397419355
7	2017	6	2.258333333	56.71	35.81667	24.786	4.618333333
8	2017	7	3.932903226	57.71709677	35.87032	24.49097	4.840645161
9	2017	8	6.376774194	75.31677419	31.98774	23.54129	4.273870968
10	2017	9	6.275	83.024	30.104	23.05767	3.485333333
11	2017	10	6.970645161	84.31483871	29.45065	22.79065	3.750967742
12	2017	11	6.349666667	88.08533333	27.862	21.132	2.816333333
13	2017	12	1.059032258	83.15225806	27.39129	17.9329	2.795806452
14	2018	1	0.06483871	74.05096774	28.99452	16.52581	4.585483871

Figure 2: Climate Dataset

Soil properties

Soil properties from different villages in Vellore collected by farmers for a period of 3 years is used as one of the set of attributes in this study.The properties include electrical conductivity, pH, presence of elements like Nitrogen, Phosphorous, Potassium, Zinc, Copper, Iron and Manganese, and the Lime status of the soil. All of these factors affect the texture of soil and in turn affect the crop productivity for that area. Table 2 display the information of each of the soil property considered in the final data set as shown in Figure 3.

	YEAR	MO	EC	pH	N	P	K	Zn	Cu	Fe	Mn	Lime Statu
2	2017	1	0.651	8.13	112	12.97	167	4.66	3.41	2.63	1.83	Profused
3	2017	2	0.55	7.35	118.5	12.71	111.3	1.339	3.068	2.809	1.969	Profused
4	2017	3	0.184	7.36	105.4	7.58	95.7	0.765	1.681	5.731	1.901	Medium
5	2017	4	0.158	7.54	121	2.01	353.9	0.758	1.867	6.602	1.699	Nil
6	2017	5	0.424	7.97	112.75	11.41	145.5	0.88	3.543	2.26	1.51	Profused
7	2017	6	0.21	7.576	77.9	5.85	133.8	0.81	0.831	4.55	2.3606	Nil
8	2017	7	0.149	7.24	104.5	6.8	140	0.992	2.48	4.214	2.925	Nil
9	2017	8	0.225	7.83	145.25	4.45	136.3	0.894	2.545	6.384	1.96	Profused
10	2017	9	0.417	7.21	103.9	6.8	154	0.7	1.773	6.472	1.569	Medium
11	2017	10	0.208	7.66	81.4	10.65	187.5	0.966	1.161	7.045	3.025	Nil
12	2017	11	0.378	7.21	124	7.98	108.2	2.88	2.28	6.02	4.55	Nil
13	2017	12	0.55	7.35	118.5	12.71	111.3	1.339	3.068	2.809	1.969	Profused
14	2018	1	0.651	8.13	112	12.97	167	4.66	3.41	2.63	1.83	Profused

Figure 3: Soil Properties Dataset

3.3. Minimum noise fraction (MNF)

The objective of MNF is to reduce the noise present in the recorded raster images for the said region. The MNF algorithm has two consecutive data reduction operations. First reduction operation is done as an estimate of the noise on the basis of the correlation matrix. Decorrelation is done and the noise is rescaled with the help of variance. Second operation creates a set of weighted components about the variance so that variance is maximized. The dominant components are used to transform the data to the original multi-spectral image. As discussed in the previous section, the components are sorted in descending order of SNR by MNF based on the image quality. Due to this we can restrict the amount of noise in the reconstructed image below a threshold level by eliminating some components from the bands with a more SNR ratio. The result of the first transformation is zero band to band correlation and noise has a unit variance and several new bands with concentrated information are created during the second transformation of MNF that is a traditional PCA [12].

The MNF Transformation on the Landsat image will be producing 2 different statistics files – the MNF statistics and the MNF noise statistics, unlike the principal component analysis.

The algorithm of Minimal noise fraction is implemented using Python. The rasterio library provided by python is used to import TIFF files. Every raster image consists of 3 components, height, width and bands. These are extracted from the image into an array to perform further transformations.

The pysptools. noise library is used to whiten the MSI cube. The covariance matrix of noise is used to rescale and de-correlate the noise in the data. This whitened image is reshaped and passed to perform Principal component analysis for dimensionality reduction. The image is reshaped for compatibility and is returned for further fitting. The cumulative sum of variance associated with the corresponding eigen vectors is also calculated.

Finally, the generated model is used to fit the original image which gives a transformed image with reduced dimensionality and variance. While applying MNF on the TIFF files, the parameters that need to be passed include the number of components present in the raster file and the actual TIFF file location. The output generated is a TIFF file which is used further for feature extraction and terrain analysis.

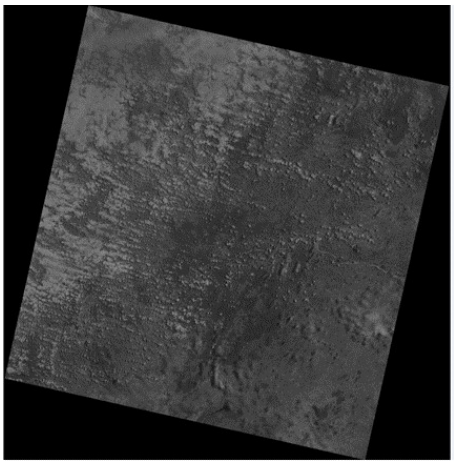

Figure 4: Original band 6 image for July-2018

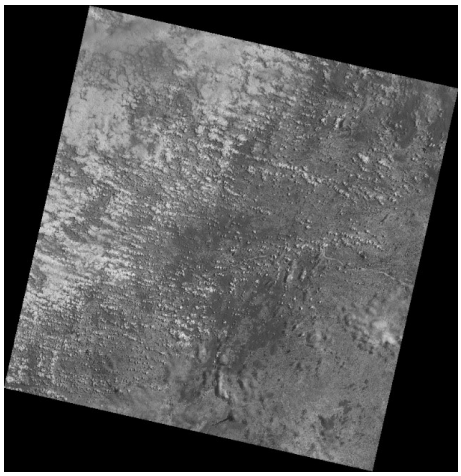

Figure 5: MNF transformed image

From the Figures 4 and 5 It can clearly be seen that after MNF transformations some of the terrain areas are cleaner and properly visible. This procedure is repeated for images of the years 2017, 2018 and 2019. MNF is performed for bands 4,5,6 and 7 since these bands will be used further for feature extraction and analysis. Figure 6 shows the Merged 6 and 7 bands for further feature extraction.

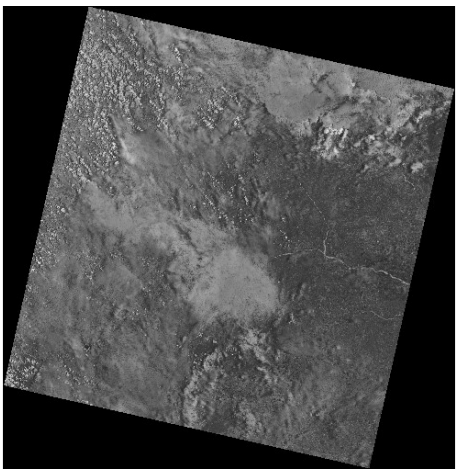

Figure 6: Merged Band6 Band7 Images

3.4. Terrain Properties

After the noise removal process, these cleaned images were used to extract terrain features. According to Landsat8 band information, bands 6 and 7 are utilized for geology and differentiating dry earth from wet earth. Rock structures, soil and other terrain variables that cannot be identified in other bands contain strong contrasts in SWIR band. These raster images were merged using QGis and various terrain features were extracted as well as derived from the merged raster.

The terrain features that were extracted using QGis include:

1. Relative Elevation (HT): Vertical distance above sea level. Calculated using the formula:

$$Relative\ Elevation = Maximum\ Curvature - Minimum\ Curvature$$

2. Slope (β): Inclination of land surface from the horizontal.
3. Aspect (A): Direction the slope faces.
4. Topographic wetness index (TWI): Ratio of local catchment area to slope.
5. Stream power index (SPI): Product of natural log of slope and flow accumulation
6. Plan curvature (C_h): Horizontal curvature.
7. Profile curvature (C_v): Vertical curvature.
8. Terrain ruggedness index (M): Macroscopic information on the surface.
9. Topographic position index (TPI): Euclidean distance to the nearest valley.

These attributes were calculated for 3 years of raster data and stored in a CSV file as shown in Figure 7.

	A	B	I	J	K	L	M	N	O	P	Q
1	YEAR	MO	Relative Elevation(Ht)	Slope(B)	Aspect(A)	Topographic wetness index(TWI)	Stream power index(SPI)	Plan curvature	Profile curvatutre	Terrain ruggedness index(M)	Topographic position index
2	2017	1	0.000209763	0.091385	190.2342	7.76221916	337.9323201	-0.000437974	1.74E-06	0.041734705	0.000796509
3	2017	2	0.000217094	0.196224	181.7789	0.523421326	258.909835	-7.42E-05	-1.15E-06	0.100073238	-0.000764993
4	2017	3	0.000212391	0.226473	151.9737	7.186380032	441.2661848	-3.89009579	0.00326681	0.099557718	-0.001125092
5	2017	4	0.000212103	0.186483	181.1155	7.039692038	540.785901	-0.000705503	1.95E-06	0.09633476	-0.000525077
6	2017	5	7.211370526	54.45706	204.0496	7.649905385	-5557.002176	0.035703509	1.76E-05	0.077895036	-0.002211621
7	2017	6	4.12E-05	0.051708	178.9633	8.158454422	213.8441832	-0.000279897	9.59E-07	0.023187639	-0.001429787
8	2017	7	0.000257711	54.76829	203.3037	7.673707847	-5798.674622	0.034796883	1.44E-05	0.088817469	-7.02E-05
9	2017	8	0.000192307	0.047217	179.6482	7.140122718	12.43981729	7.98E-05	-2.91E-06	0.065100092	0.001181284
10	2017	9	0.000164621	0.101625	184.0579	8.157445724	455.8988538	0.000360116	-3.22E-06	0.047729803	-0.003298848
11	2017	10	0.000120329	54.08588	184.7285	2.67456194	-10.74733294	-0.068381719	-0.022380604	15.83928045	0.014324953
12	2017	11	0.000157184	0.109093	184.3187	4.970736285	4.61E-08	-0.000453528	-1.23E-06	6.04E-08	8.09E-11
13	2017	12	7.69E-05	0.145682	181.6255	7.254822567	390.736034	0.00170944	-8.80E-06	0.114594222	-0.000137818
14	2018	1	4.12E-05	0.004653	150.7584	9.700803335	7.155187514	-5.16037715	0.003904236	0.121454009	-0.002379274
15	2018	2	0.000209763	0.212725	181.9436	7.156342571	613.313396	0.00048957	-7.73E-06	0.102930713	-0.000601496
16	2018	3	0.000257711	0.233158	181.3752	6.902678252	763.1367331	0.000136217	-3.38E-06	0.118619396	-0.000768436

Figure 7: Final input data csv

3.5. Normalized Difference Vegetation Index

Plants absorb solar radiations in the spectral region, which they use as a source of energy in the process of photosynthesis. The chlorophyll present in the leaves absorbs visible light (Band 4) for photosynthesis. However, the cells of the leaves, strongly reflect near-infrared light (Band 5). These wavelengths of light are affected by the amount of leaves a plant has. Normalized Difference Vegetation Index (NDVI) is a standardized way to measure healthy vegetation. It normalizes reflection in the Near Infra-red wavelength and absorption in the red wavelength. It used the band 5 (NIR) and band 4 (red) channels in its formula. Vegetation strongly reflects NIR and absorbs red light. NDVI measures the difference between these factors. It generally ranges from -1 to +1. When it has a low value, it means there is no or less vegetation in that area.

The formula for calculation of NDVI can be given as:

$$NDVI = \frac{(NIR - Red)}{(NIR + Red)} \tag{3.1}$$

QGIS is a geographic information system application that supports raster calculations as shown in Figure 8. It is used because of its fast processing time and better rendering capabilities. The value for NDVI can be calculated using the raster calculator provided in QGIS. The pre-processed images for bands 4 and 5 are loaded into the application. The raster calculator does all the calculations for NDVI using the formula provided by the user.

The raster images are all grayscale images, which does not convey the vegetation properties very well. However, once the NDVI calculations are complete the colour pallet can be changed to a relevant colour, such as green. The output of the raster calculations gives a raster image consisting of range of colour shades. The darker areas imply there is a presence of healthy vegetation and the whiter areas imply there is little or no vegetation. This can help in analysing the properties of those areas where healthy vegetation is present as seen in Figure 8. The statistical values for NDVI can be taken for further processing. These are recorded for each month for a span of three years.

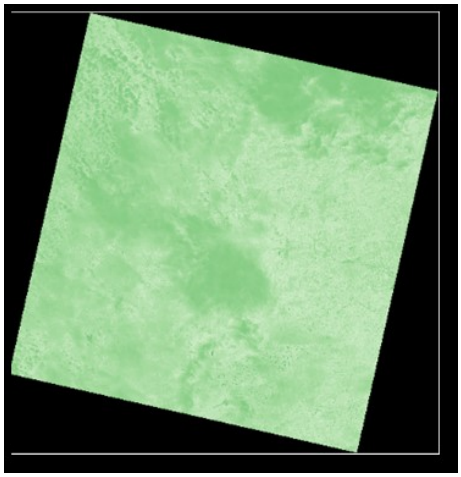

Figure 8: Calculating NDVI in QGis

4. CLASSIFICATION MODEL (ENSEMBLE MODELLING)

4.1. Logistic Regression

Logistic regression is another widely used classification method when the target variables are categorical. The output is in the form of probability. It gives the probability of a point with respect to the class it belongs to. The mathematical model for the regression line is given by,

$$\frac{p}{1-p} = e^{y} \tag{4.1}$$

Where,
y: given sample
p: probability that the sample belongs to the given class.
Input values are combined using weights or coefficient values to predict an output value.

4.2. Decision Tree Classifier

The Decision Tree Classifier is a method in which data is continuously split based on a certain parameter. In this study the target variable is categorical or discrete, hence the model builds a classification tree using a process called binary recursive partitioning. This is an iterative method that splits the data into partitions recursively.

The main idea is to split the data into clusters and create an associated decision tree simultaneously. At the end of this process a final tree that covers the whole training dataset is returned. The data is split on the feature that provides the highest information gain. In simple terms, information gain is about finding the attribute that returns the most homogeneous branches.

This splitting process must be ideally performed till the samples at each leaf node belong to the same class. Although, one of the disadvantages of using decision trees is its tendency to overfit the model, this can be overcome by setting a limit on the depth of the tree. However, the final leaves may not all be homogeneous.

4.3. Support Vector Machine

The Support Vector Machine (SVM) is a classification technique which determines the statistical data of spectral images and other climatic and soil property attributes. SVM is a non-probabilistic linear binary classifier which identifies the feasible hyper to separate two classes in a high-dimensional space. It also considers the data points that are located on the edge of the class distributions. These are called as support vectors.

SVM is initially used for classification of soil status of an area based on terrain and other types of statistical information derived from spectral images. The use of Kernel functions helps in plotting the input data into a hyperspace. This area is where the separations are performed. The ε-insensitive loss function plays a major role to get an optimal hyperspace for fitting the data and further for predictions. This function tolerates minimal errors than the constant ε set as a threshold.

The climate attributes, terrain features, NDVI and soil properties are all used together to predict the soil status of the area under study. Prediction is done for each month throughout the three-year span. The ratio of training and testing data plays and important role in the generation of the model. If the training data is not sufficient it might lead to incorrect and inefficient classification and prediction. If the testing data is not enough for the model to analyse and compare it affects the performance metrics of the classifier. For this study 80% of the data is used for training the SVM classifier and 20% is used to test the predictions. These predictions can be compared with the actual target variable of the test data for model evaluation.

For this study the support vector machine classifier is implemented using python's Scikit learn library. The SVMs in scikit-learn support both sparse and dense sample vectors as input. It is effective in high dimensional spaces and is versatile.

4.4. Ensemble method - Voting Classifier

Ensemble modelling is a machine learning technique the combines several base models to produce optimal results. Rather than relying on one model, ensemble methods takes into account the prediction of each model to give the final prediction based on the sample models. In this paper, the Voting Classifier is used to train the ensemble model. The voting classifier trains on an ensemble of the three models described above. The output is predicted based on the highest probability of the chosen output class. The predictions of each of the three classifiers are aggregated and passed on to the voting classifier. The type of voting used in this paper is soft voting, which means the output is the average probability of that class. The results generated will also be more precise than the regular classification techniques such as Decision Trees, Logistic Regression and Support Vector Machines.

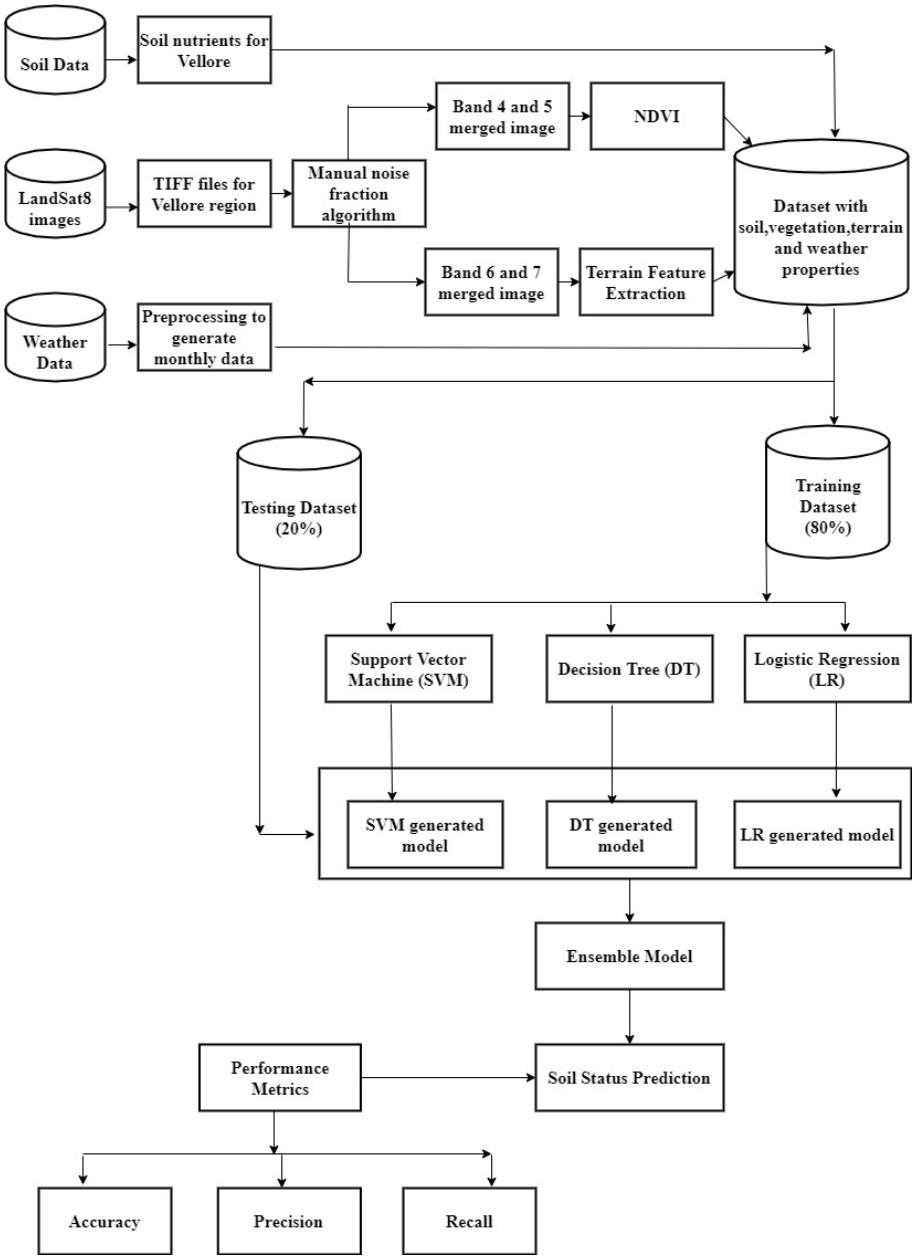

Figure 9: Workflow

5. RESULTS

The performance metrics of classifiers include confusion matrix, accuracy of the model, precision of the model and also the F1 score and recall of the classifier. The confusion matrix is count of the negative values identified correctly, i.e., true negatives, the positive values that were correctly identified by the model, i.e., true positives, the false negatives and the false positives. The output generated comes in the form of a 2×2 matrix where $C_{0,0}$ is the true positives, $C_{1,0}$ is the false negatives, $C_{1,1}$ is the true negatives and $C_{0,1}$ is the false positives. The accuracy of the model is the percentage of records correctly classified. It is given by the formula,

$$Accuracy = \frac{True\ Positives\ + True\ negatives}{True\ Positives\ + False\ positives + True\ negatives + Fulse\ negatives} \tag{5.1}$$

The precision of the classifier id the percentage of only the predicted positives which are actually true. It is given by,

$$Precision = \frac{True\ Positives}{True\ Positives\ + False\ positives} \tag{5.2}$$

Recall is about capturing the cases that have been classified correctly. It gives the proportion of the cases that are actually true, given by the formula,

$$Recall = \frac{True\ Positives}{True\ Positives\ + Fulse\ negatives} \tag{5.3}$$

F1-score represents both precision and recall. It takes the harmonic mean of the two and is closer to the smaller metric. It is given by,

$$F1 - score = \frac{2 * precision * recall}{precision + recall} \tag{5.4}$$

The satellite images after removal of noise using MNF and extraction of suitable features through various methods are classified using the SVM model and their accuracies are compared through different performance metrics and by analysing the confusion matrix.

Three classes of Soil Textures were taken into consideration for the Vellore region Clay Loam, Sandy Clay Loam, Sandy Loam as the target variables for this paper. The weather conditions like Precipitation, Humidity, Atmospheric Pressure etc., were taken into consideration, NDVI calculated from the Landsat 8 images and the terrain properties like slope of the land, aspect, Terrain ruggedness, Elevation, Curvature etc., were the factors considered. With ensemble voting classification model the predication accuracy for each model can be compared as seen the table 1.

Table 1: Accuracy Comparison

Accuracy	Model
0.60 (+/- 0.25)	Logistic Regression
0.87 (+/- 0.19)	Decision Tree
0.50 (+/- 0.00)	Support Vector Machine
0.90 (+/- 0.08)	Ensemble

The change in Soil Texture along the course of 3 years where the values 1 represent clay loam, 2 represent Sandy Clay Loam and 3 is Sandy Loam is graphed in Figure 11. A significant increase in accuracy is seen with reduced variance values when all the three classifiers are stacked for the Voting Classifier ensemble model. The 80% of the Data was taken to be Training data and 20% was considered to be Testing data. The confusion matrix result is plotted in Figure 12 and classification report is displayed in table 2.

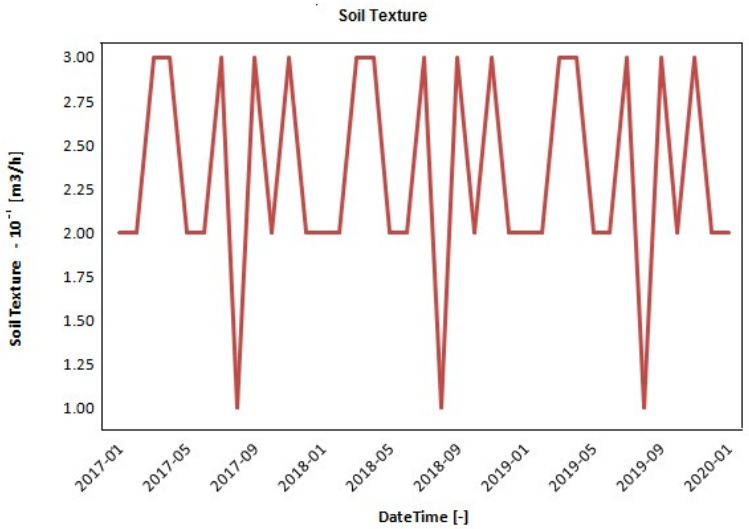

Figure 10: Changes in Soil Texture with Time

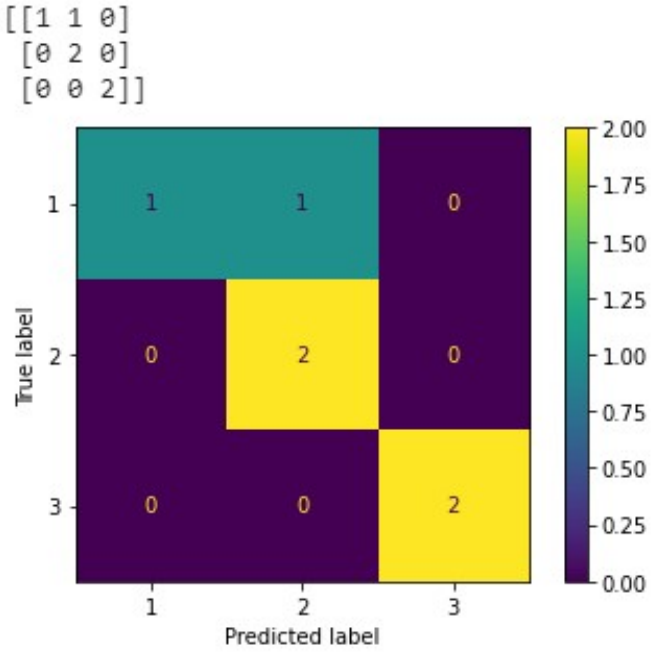

Figure 11: Confusion Matrix

Table 2: Performance Metrics of the model

No.	Precision	Recall	F1-score	Support
1	1.00	0.50	0.67	2
2	0.67	1.00	0.80	2
3	1.00	1.00	1.00	2
Macro avg accuracy	0.89	0.83	0.82	6
Weighted avg accuracy	0.89	0.83	0.82	6

The consolidated performance metrics are as follows:

1. Precision: 67%
2. Recall: 100%
3. F1-Score: 80%
4. Accuracy: 83%

6. CONCLUSION

Identifying the soil texture and other properties can be automated by deploying machine learning techniques. The noise of images captured in Landsat 8 is removed by MNF. Three different classifiers namely logistic regression, support vector machine and decision tree are integrated and voting is performed for each test data. The cross-validation accuracy for the model determined is 94.44% . The proposed model is accurately able to classify the soil texture on the basis of Landsat 8 images, weather conditions and the features extracted from it with the soil mineral content data.

References

[1] G. Luo, G. Chen, L. Tian, K. Qin, S.E. Qian, Minimum noise fraction versus principal component analysis as a preprocessing step for hyperspectral imagery denoising. Canadian Journal of Remote Sensing, 42.2 (2016): 106-116.
[2] O.R. Seryasat, O. Rahmani, J. Haddadnia, Evaluation of a new ensemble learning framework for mass classification in mammograms, Clinical breast cancer 18.3 (2018): e407-e420.
[3] O.R. Seryasat, H.G. Zadeh, M. Ghane, Z. Abooalizadeh, A. Taherkhani, F. Maleki, Fault Diagnosis of Ball-bearings Using Principal Component Analysis and Support-Vector Machine. Life Science Journal, 10.1 (2013): 393-397.
[4] J. Haddadnia, O.R. Seryasat, H. Rabiee, Thyroid Diseases Diagnosis Using Probabilistic Neural Network and Principal Component Analysis. Journal of Basic and Applied Science Research, 3.2 (2013): 593-598.
[5] F. Frassy, G. Dalla Via, P. Maianti, A. Marchesi, F.R. Nodari, M. Gianinetto, Minimum noise fraction transform for improving the classification of airborne hyperspectral data: Two case studies. In 2013 5th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), IEEE (2013, June): 1-4.
[6] B. Dhruv, N. Mittal, M. Modi, Analysis of different filters for noise reduction in images, In 2017 Recent Developments in Control, Automation & Power Engineering (RDCAPE), IEEE (2017): 410-415.
[7] P.A. de Oliveira Morais, D.M. de Souza, M.T. de Melo Carvalho, B.E. Madari, A.E. de Oliveira, Predicting soil texture using image analysis. Microchemical Journal, 146 (2019): 455-463.
[8] H. Ren, C. Du, R. Liu, Q. Qin, G. Yan, Z.L. Li, J. Meng, Noise evaluation of early images for Landsat 8 Operational Land Imager. Optics express, 22.22: (2014): 27270-27280.
[9] L. Gao, B. Zhang, X. Sun, S. Li, Q. Du, C. Wu, Optimized maximum noise fraction for dimensionality reduction of Chinese HJ-1A hyperspectral data. EURASIP Journal on Advances in Signal Processing, 2013.1 (2013): 65.
[10] https://earthexplorer.usgs.gov/.
[11] https://power.larc.nasa.gov/data-access-viewer/
[12] H. Yang, X. Zhang, M. Xu, S. Shao, X. Wang, W. Liu, H. Liu, Hyper-temporal remote sensing data in bare soil period and terrain attributes for digital soil mapping in the Black soil regions of China. Catena, 184 (2020): 104259.
[13] X. Liu, B. Zhang, L. Gao, D. Chen, A maximum noise fraction transform with improved noise estimation for hyperspectral images. Science in China Series F: Information Sciences, 52.9 (2009): 1578-1587.
[14] V. Yogesh, K. Yogendra, Removal of Salt and Pepper Noise from Satellite Images. International Journal of Engineering Research & Technology (IJERT), 2 (2013): 2051-2058.