

# Feature-Based Image Retrieval (FBIR) System for Satellite Image Quality Assessment Using Big Data Analytical Technique

Saurabh Srivastava<sup>1</sup>, Tasneem Ahmed<sup>2</sup>

<sup>1,2</sup> Advanced Computing and Research Laboratory, Department of Computer Application, Integral University, Lucknow, Uttar Pradesh, India

Email: <sup>2</sup>tasneemfca@iul.ac.in

## ABSTRACT

The earth observatory sources generate various types of satellite images with different resolutions such as spatial, spectral, and temporal resolution. Satellite images are very expensive but also there are some freely available satellite images which are used in various applications such as land use land cover classification, agriculture monitoring, fire monitoring, urban monitoring, and flood monitoring, etc. A single satellite generates several terabytes of data per day which means a single satellite creates a huge amount of data to be analyzed. Traditional analysis techniques are not suitable for satellite images because it is less capable to handle large and complex datasets, but big data analysis techniques can be useful to extract meaningful information from the satellite images and it can apply on real-time data as well as offline data. Some of the challenges may occur during the implementation of satellite images with big data analysis such as modeling, processing, mining, querying, and distributing large scales of repositories because it refers to typical datasets which may generate problem during capture, storage, analysis, locating, identifying and securing, and understanding of the data. In this paper, a Feature-Based Image Retrieval (FBIR) system is being developed which will always serve those satellite images which can produce a better result for interpretation and analysis of the earth surface monitoring. The system will always allow retrieving the best possible available image (i.e. less affected image in terms of weather conditions and cloud cover) to the end-user for future use.

## Keywords

Similarity measurement, feature extraction, image classification, SVM, big data

Article Received: 10 August 2020, Revised: 25 October 2020, Accepted: 18 November 2020

## Introduction

The satellite images are the images of earth or other planets acquired by the various satellites, which are operated by Government agencies and business organizations around the world. The satellite image provides a standard approach to earth observation and geospatial knowledge. Various sources such as National Aeronautics and Space Administration (NASA), United States Geological Survey (USGS), and National Remote Sensing Centre (NRSC) are providing satellite images with high-quality spatial, spectral, and temporal resolutions. Satellite Image is an important source for covering the information of manmade and natural resources, and it has been used in various applications such as land use and land cover mapping, change detection, vegetation monitoring, and time series analysis, etc. [1]. Currently, more than 200 satellite sensors are capturing multi spatial and multi-temporal satellite images through multiple sensors available in the orbit [2], and providing daily, weekly or monthly basis images for interpretation and analysis of earth or other planets surface monitoring. The advanced sensors and earth observation devices are providing high resolution of satellite images,

the datasets of satellite images contain multi-dimensional and complex structured metadata so it is difficult to load and reside the massive data in large memory [2]. The analysis of an image is a complicated and time-consuming process, especially when the image shares multispectral data. A single satellite generates several terabytes of data every day, and it is difficult to process each and every image to extract the meaningful information from them [2].

Satellite sensors are producing a large number of satellite images on a daily, weekly, and monthly basis. NASA and the Earth Observation System Data and Information System (EOSDIS) have successfully managed satellite image data that was expected to be from 7.5 PB during the year 2012 [2]. Earth Science Data and Information Systems (EDIS) had 7.5 PB data with approximately 7000 unique datasets and 1.5 million users in 2013 [3]. EOSDIS has already distributed more than 4.5 million GB of data but the understanding growth of EOSDIS data is 4 TB per day [2]. IBM explained about the current era of data that more than 2.5 quintals bytes of data are being generated every day. In other words, "90% of the data in the world today has been created in the last two years

alone" [3]. After some time, the produced data will require a large volume of space to store, and stored data will act as big data [4] which arises from communication technology and the rapid development of computer applications but the origins of big data creation are data collection, data analysis, and data management [5]. Data collection define that various satellite sensors are providing different resolutions of satellite images per day, and stored in a single repository. Data size is increasing worldwide, and traditional analysis techniques are less capable to handle large and complex datasets. Also, these methods are not designed for large quantities and complex datasets, so these data analysis methods are not suitable for computing large and complex datasets. Statistics and machine learning techniques are widely used to extract meaningful information from large and complex datasets, and data management define that distributed file system and cloud computing are used for big data management [6]. Big data analysis is a modern approach that reduces the challenges of traditional analysis approaches [7].

Big data has started to significantly influence global production, circulation, distribution, and consumption patterns. It is changing humankind's production methods, lifestyles, mechanisms of economic operation, and country governance models. Big data occupies a strategic high ground in the era of the knowledge-driven economy and it is a new strategic resource for all nations [8]. Satellite sensors are smart and intelligent devices and generate a large amount of data. After some time, generated data take a large volume and act as big data. The big data repository holds structured, unstructured, and semi-structured data which involves additional processing [4].

The size of the data all over the world is increasing day by day while working with satellite images the advancement in technology is needed [7]. Big data analysis includes aggregation, analysis, and the storage of large volume data. Its analysis considers the data quality and data normalization to take place and data is molded into rows and columns. The molded data is consigned as an enterprise data warehouse for data analysis [6]. Big data analytics is a modern approach to extract meaningful information from the large and complex dataset. Various big data analytical techniques are available, some

commonly used big data analytical techniques are ensemble data analysis, association rule analysis, machine learning, precision Analysis, divide and conquer analysis [5]. Big data and its analytics techniques contribute to many areas such as the public sector, learning, industrial and natural resources, transportation, banking zones, and fraud detection etc. [9], but in this literature main focus is on the application which are using satellite images. Many prediction analysis applications are available which use real-time satellite images for better interpretation of satellite images for land surface monitoring. In brief literature review, it is observed that many researchers have used various big data techniques with satellite images such as Trinity framework, RealBDA architecture, Remote sensing big data analytics architecture, and Big data image processing framework [3,10,11,12]. Trinity framework is proposed for a better understanding of data in terms of satellite image applications. The trinity framework defines some common challenges such as data computing, data collaboration, data methodology and some individual challenges such as data transport, data storage, data delivery, data representation, data fusion, data visualization, data identification, data interpretation, data deployment from different perspectives such as forest planning, land development and land use urban planning management, and cross assessment and yield forecasting [3]. Trinity architecture classified big data by using three different perspectives related to who owns big data, who has innovated big data methods and methodology, and who needs big data application [3]. The RealBDA Architecture can process real-time big data generated by the satellites. RealBDA architecture is developed to reduce the complexity of big data storage and analysis efficiently [10]. The earth observatory satellite provides lots of data that is called raw data, the received data consist of a wide variety and the identification of hidden parameters in received data is too difficult, this type of problem can be solved by RealBDA architecture. The RealBDA architecture for satellite image applications consists the three major units such as data pre-processing unit, data analysis unit, and data post-processing unit [10]. The remote sensing big data analytics architecture for satellite images is capable of dividing load balancing and parallel processing of the satellite images and also store

incoming raw data to perform the analysis. Hadoop plays an important role in this architecture and the architecture can analyze real-time data as well as offline data [11]. The big data analytics architecture collects the satellite image data from different sources and extracts useful information from it, and transmits it to the earth base station for further data processing [11]. The developed architecture and algorithms have been implemented in Hadoop with map-reduce. The Big Data analytical architecture for satellite image applications consists the three major units such as remote sensing big data acquisition unit (RSDU), data processing unit (DPU), and data analysis and decision unit (DADU) [11]. The objective of Big data image processing framework is to analyze a large number of satellite images using multiple nodes. The architecture takes images as a job and it produces efficient robust and low latency solutions for a large number of satellite images and analyzes the satellite images in a distributed way. The proposed architecture has the name and data nodes, which are connected through Wi-Fi [12]. Name node divides the image into multiple parts and sends it to all available data nodes, the data nodes process and store it in local memory. Name node contains metadata in their memory which holds the information of the data node and each part of the image. If any user wants the result then the user can directly access the data node and get the result [12].

The challenges that may occur during the implementation of satellite images with big data analysis can be categorized into two categories such as common challenges and individual challenges. The common challenges include big data computing, big data collaboration, and big data methodology. Where, big data computing designed for high-performance systems that are more heterogeneous and capable to integrate resources in a different location and develop a distributed system for data collection. Big data collaboration is an important challenge for a government institution to share data unless all participants can achieve material benefits and incentives in data sharing that out weight the task for instance, even if NASA is now sharing a significant amount of satellite images data under the open government initiative [3]. Big data methodology is designed to notify big data problems. The main objective of choosing or

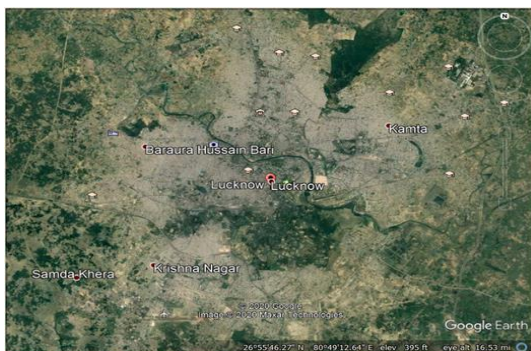
designing a methodology of big data needs a mechanism that can tackle problems that may occur during an analysis of satellite images such as data preparation, information extraction, data modeling, etc. There are some individual challenges such as data identification challenges, big data possession, data deployment, data representation, data fusion, data visualization, and interpretation with respect to the satellite image processing in the perspective of satellite image processing [3].

Earth observation technology provides a new vision and methods for earth science research. Big data is a radical modernization that allows the development of new methods in scientific research. A satellite image is at the forefront of the integration of geoscience, information science, space science, and technology; and satellite data may provide new prospects for the development and research in the earth science domain [8]. Various earth observatory sources providing satellite images every day, some of these are affected due to weather conditions and cloud cover. In this study, big data analytics and satellite images have been used together and developed feature-based image retrieval (FBIR) system that allows retrieving those images which can produce an efficient result for interpretation and analysis of earth surface monitoring, and also suggests some similar images of retrieved image for future use. Various satellite image retrieval system exists that are providing satellite images for interpretation and analysis of earth surface monitoring. Every image retrieval system has an image repository that contains a variety of image but some images in the repository may be affected due to weather conditions and cloud covers, and the affected images are not suitable for post-processing because it cannot produce an efficient result for interpretation and analysis of earth surface monitoring. The traditional satellite image retrieval system is less capable to indicate the quality of satellite images before image retrieval. The proposed FBIR system is capable to indicate the quality of satellite images before the image retrieval and provide the facility to download those images which can produce an efficient result for interpretation and analysis of earth surface monitoring.

## Study Area and Satellite Data Used

### Study Area

The study area is located between latitude 26.945746° to 26.734630° and longitude between 80.74451° to 81.142585°. It covers Lucknow city and its nearby areas. Lucknow, a large city in northern India, and is the capital of the state of Uttar Pradesh. Toward its center is Rumi Darwaza, a Mughal gateway. Nearby, the 18th-century Bara Imambara shrine has a huge arched hall. Upstairs, Bhool Bhulaiya is a maze of narrow tunnels with city views from its upper balconies. Close by, the grand Victorian Husainabad Clock Tower was built as a victory column in 1881. Figure 1 shows the study area on Google Map. The study area has been chosen for its varied landscapes: water (source: Gomti river, the most of which remains dry in summers), urban (source: Lucknow city), tall vegetation (source: dense tree cover in city Malihabad, Kakori), bare soil, short vegetation (source: cropland and grassland).



### Satellite Data Used

The most commonly used optical satellite images are Landsat images which have been used for various applications such as agriculture monitoring, geology, forestry, regional planning, land cover, and land classification, etc. The first satellite of the Landsat series launched in 1972 [13]. In this study, Landsat-8 images are undertaken which is the latest and eighth satellite of the Landsat series launched on February 11, 2013, and the seventh satellite that successfully reached orbit. Landsat-8 is the product of USGS EROS Centre, and it holds an operational land imager (OLI) and thermal infrared sensor (TIRS) with 11 multi-spectral bands. The Ultra blue (Coastal/Aerosol), Blue, Green, Red, Near-Infrared (NIR), 2-Shortwave Infrared (SWIR), Cirrus bands are with 30 meter spatial resolution, Panchromatic band is with 15 meter, while 2-Thermal Infrared (TIRS) bands are available with 100 meter spatial resolution and the revisit time (temporal resolution) of Landsat -8 is 16 days [14]. To create a big data repository, approximately 140 Landsat 8 OLI raw satellite images have been downloaded. Some of the details of Landsat images such as acquisition ID and acquisition dates are shown in table 1.

Figure 1. Google Map Image of Lucknow City.

Table 1. Some of Satellite Images Details

<i>Acquisition ID</i>	<i>Acquisition Date</i>	<i>Image ID</i>
LC08_L1TP_144041_20161217_20180524_01_T1	2/17/2016	IMG_1
LC08_L1TP_144041_20160405_20170327_01_T1	4/5/2016	IMG_2
LC08_L1TP_144041_20160217_20170329_01_T1	2/17/2017	IMG_3
LC08_L1TP_144041_20170408_20170414_01_T1	4/8/2017	IMG_4
LC08_L1TP_144041_20170424_20170502_01_T1	4/24/2017	IMG_5
LC08_L1TP_144041_20180121_20180206_01_T1	1/21/2018	IMG_6
LC08_L1TP_144041_20181223_20181227_01_T1	12/23/2018	IMG_7
LC08_L1TP_144041_20190209_20190221_01_T1	2/9/2019	IMG_8
LC08_L1TP_144041_20190313_20190313_01_RT	3/13/2019	IMG_9
LC08_L1TP_144041_20190329_20190404_01_T1	3/29/2019	IMG_10
LC08_L1TP_144041_20191210_20191217_01_T1	12/10/2019	IMG_11
LC08_L1TP_144041_20200127_20200210_01_T1	1/27/2020	IMG_12
LC08_L1TP_144041_20200212_20200225_01_T1	2/12/2020	IMG_13
LC08_L1TP_144041_20200416_20200416_01_RT	4/16/2020	IMG_14



## Theoretical Background

### Pre-Processing of Satellite Images and Calculation of Vegetation Indices

**Satellite Image Pre-Processing:** Satellite Image Pre-processing: The downloaded Landsat 8 images are pre-processed to convert the digital numbers into the reflectance values by applying the radiometric calibration in ENVI 5.1. The radiometric calibration process involves converting the raw signal from a detector into the expected aperture spectral radiance/reflectance. The raw signal is expressed as 12-bit digital numbers as a result of the focal plane electronics [15]. A large part of the success of the satellite program can be attributed to the knowledge of the radiometric properties of the satellite sensors. The radiometric calibration helps to characterize the operation of the satellite sensors, but, also allows the full satellite data set to be used in a quantitative sense for such applications as land use and land-cover change, change detection, agriculture monitoring etc. [16].

Radiometric calibration has been applied on each and every image to produce the true color composite (TCC) image, which is generated by using the red, green, blue bands and it has given contrast signature for anorthosite is a mixture of yellow and white color. TCC is one of the good combinations for anorthosite discrimination as it represents the true color of an object as seen through the naked eye [17]. The true color composite image contains RGB (where R=Red, G=Green, and B=Blue) Bands, which shows the shapes and sizes of the various features. The shape defines the geometric outline of an object, and this outline gives information about the nature and geometry of the object. Size defines the magnitude of an object or a single dimension of the object (e.g., the length of a river) [18]. TCC images are used to extract the various features. Figure 2 shows the true color composite (TCC) image of the study area.



**Figure 2. TCC image of Lucknow City.**

**Feature Extraction:** Image feature is a piece of information about the content of an image, and it is categorized by color, texture, and shape. The color feature is the most popular feature of an image and it can be effortlessly obtained from pixel intensities. It is used to describe color histograms over the whole image and calculated by the average and standard deviation of the color intensity of each color component [19]. The texture feature is a low-level feature of an image, and it is used to describe the detail of the spatial distribution of different patterns in an image such as the spatial arrangement of color and intensities in an image [20]. Shape features are less developed to compare than color and texture features of an image because of their inherent complexity of representing shapes. It is used to describe the shape of the different regions present in an image such as external boundary, outline, and an external surface [20]. The feature extraction algorithm is used to extract the spatial objects of interest, and the spatial information defines the different spatial features of the image. Satellite image holds various spatial features such as color, texture, a region of interest, and shapes. In this study, the color moment feature has been extracted which is the most popular algorithm for color feature extraction. The color moment algorithm extracts the mean, standard deviation, and skewness [21]. In this study, the mean and standard deviation of the red, green, and blue bands of an image have been used. In the image, the mean can be understood as the average color value, and the square root of the variance distribution can be defined as the standard deviation. The formula of mean and standard deviation are given below [21].

$$E_i = \sum_{j=1}^N \frac{1}{N} P_{ij} \quad (1)$$

$$\sigma_i = \sqrt{\left(\frac{1}{N} \sum_{j=1}^N (P_{ij} - E_i)^2\right)} \quad (2)$$

Where 'i' is the current channel index, 'j' is the number of channels.  $E_i$  is the first moments (mean) of the two image distributions and  $\sigma_i$  is the second moments (standard deviation) of the two image distributions.

**Image Classification:** Image classification is a technique to group the pixels that have similar digital number values across the series of image band information. Image classification is now driven by many statistical learning techniques [22]. The satellite image classification is responsible to group the different targets into corresponding categories according to their characteristics, which are used for various applications such as crop classification, land cover classification, and post-classification change detection [23]. The classification techniques are generally categorized into two categories that are supervised, and unsupervised classification [24]. In supervised classification, the samples of known informational classes (training sets) are used to classify the pixels of unknown identify, and the unsupervised classification examines the classes based on natural grouping present in the image values. The system determines spectrally separable classes and then defines their information values [24]. The most commonly used unsupervised classification techniques are K-Means and ISO-Data, while the supervised classification techniques are parallelepiped, minimum distance (MD), maximum likelihood (ML), support vector machine (SVM), and artificial neural network (ANN).

Minimum distance (MD) classification is a supervised image classification technique. In this classification, the pixels are classified using their distance from the mean spectra of the pre-defined classes. The first mean vector for each class is calculated based on the training dataset, and the MD classification is mathematically simple to other supervised classification [25]. Maximum likelihood (ML) classification is a supervised image classification technique in which the probability value of pixels is taken into consideration for classifying the pixels. In this

technique, the probability of each pixel belongs to a class is calculated. ML classifier assumed that all input bands have a normal distribution; this method is highly efficient when it comes to classifying the satellite images [25]. The support vector machine (SVM) technique of classification constructs a set of hyper-planes in a high dimensional space, which is used for regression or classification. SVM uses a non-parametric approach and can handle more input data efficiently. The accuracy and performance of this approach depend upon the hyper-plane selection [25], and the parameter of the kernel. The structure of SVM is complex to compare to other approaches, and the SVM gives low result transparency [25]. An artificial neural network classifier is a non-parametric approach to image classification, and the performance and accuracy of ANN depend on the structure of the network and inputs. An ANN classifier holds some functions of the human brain and the normal tendency for storing experimental knowledge. An ANN contains the set of sequence layer, each layer of the neural network have the number of neurons. All layers are tightly linked by weighted connections to all neurons of the processing and succeeding layer [25]. In this study, maximum likelihood (ML), support vector machine (SVM), and artificial neural network (ANN) classification techniques have been applied on various vegetation indices.

**Vegetation Indices:** Image classification is a Satellite images have been successfully used in various fields, such as classification and change detection of land use and land cover. However, satellite image processing involves a few pre-processing procedures in addition to classification and change detection [26]. In this study, vegetation indices have been used as a pre-processing procedure of image classification to highlight the desired land features of the land surface [27]. Vegetation indices are a mathematical spectral transformation formula of two or more wavelength (bands). The development of an index is based on unique patterns of each land cover and spectral response of the signature feature [27]. The vegetation indices can be categorized by spatial and spectral indices and a combination of spectral data [27]. Various image indices have been developed to categorize the built-up area, vegetation, barren

land, and water bodies. In this study, NDBI, NDVI, SAVI, and AWEI indices have been used to highlight the desired features of the land surface. The details and mathematical formulation of some of the indices are given below, which are as follow:

- Normalized difference built index uses the difference and the ratio of the Middle-infrared band and Infrared band to highlight the built-up areas and the advantage of NDBI is its simplicity and computational speed [27]. The formula of NDBI is given below:

$$\text{NDBI} = \frac{(\rho_{\text{SWIR1}} - \rho_{\text{NIR}})}{(\rho_{\text{SWIR1}} + \rho_{\text{NIR}})} \quad (3)$$

Where,  $\rho_{\text{SWIR1}}$  is the reflectance of the SWIR1 band, and  $\rho_{\text{NIR}}$  is the reflectance of the NIR band.

- Soil vegetation spectral behavior is first modeled graphically adjustment of NIR and Red wavelength space origin to the various isoline convergence point. Shifting the origin toward a negative value is equivalent to adding a constant (L) to the Red and NIR reflectance data [28]. The formula of SAVI is given below:

$$\text{SAVI} = \frac{(\rho_{\text{NIR}} - \rho_{\text{RED}}) * (1 + L)}{(\rho_{\text{NIR}} + \rho_{\text{RED}} + L)} \quad (4)$$

Where,  $\rho_{\text{NIR}}$  is the reflectance of the NIR band and  $\rho_{\text{RED}}$  is the reflectance of the RED band, L is the soil adjustment factor, and the adjustment factor can vary according to vegetation density [28]. L may within a range between 0 and 1, and 1 to 100 to analyze the effect and sensibility of NIR and Red data [28]. In this study, the soil adjustment factor  $L=0.5$  has been used to reduce the soil noise considerably throughout the range in vegetation indices.

- Automated water extraction index is useful to increase the contrast between water and other dark surfaces. The main objective of AWEI is to maximize the separability of water and non-water pixels [29]. Accordingly, two separate indices are available to perfectly suppress non-water pixels and extract water surfaces with improved accuracy [29]. The formula of AWEI is given below:

$$\text{AWEInsh} = 4 * ((\rho_{\text{GREEN}}) - (\rho_{\text{SWIR1}})) - ((0.25 * (\rho_{\text{NIR}})) + (2075 * (\rho_{\text{SWIR2}}))) \quad (5)$$

$$\text{AWEIsh} = (\rho_{\text{BLUE}}) + (2.5 * (\rho_{\text{GREEN}})) - (1.5 * ((\rho_{\text{NIR}}) - (\rho_{\text{SWIR1}}))) - (0.25 * (\rho_{\text{SWIR2}})) \quad (6)$$

Where,  $\rho_{\text{GREEN}}$  is the reflectance of GREEN band, and  $\rho_{\text{SWIR1}}$  is the reflectance of SWIR 1 band,  $\rho_{\text{NIR}}$  is the reflectance of NIR band, and  $\rho_{\text{SWIR2}}$  is the reflectance of SWIR 2 band, and  $\rho_{\text{BLUE}}$  is the reflectance of the BLUE band.

AWEInsh is an index formulated to perfectly eliminate the non-water pixels, including dark built surface in an area with built-up background and AWEIsh is mainly formulated for further improvement of accuracy by removing shadow pixels that AWEInsh does not perfectly eliminate [29]. AWEI is a simple approach, and also robust under various environmental conditions and different types of water bodies [29].

- The normalized difference vegetation index employs the multi-spectral satellite image techniques to find the spectral signature of different objects such as vegetation, land cover, urban areas [30]. NDVI is widely used as a vegetation index, the NDVI is calculated as the ratio of between measured canopy reflectance in the Red and NIR bands [30]. The formula of NDVI is given below:

$$\text{NDVI} = \frac{(\rho_{\text{NIR}} - \rho_{\text{RED}})}{(\rho_{\text{NIR}} + \rho_{\text{RED}})} \quad (7)$$

Where,  $\rho_{\text{NIR}}$  is the reflectance of the NIR band, and  $\rho_{\text{RED}}$  is the reflectance of the RED band.

By taking above mentioned indices, layer stacking has been done to perform the ML, ANN and SVM classification techniques.

### Big Data and Big Data Analytical Technique

**Big Data:** Big data refers to the group of datasets that is large and complex, the size of datasets is continues growing due to numerous intelligent devices. Big data management and implementation of the application are very difficult due to the time complexity and space complexity of data [5]. Big data holds the characteristics of multidimensional data,



currently, there are five dimensions such as volume, variety, velocity, value, and veracity. 5 V's dimensions of big data are the pillars of big data. The characterization of big data as 3V's dimensions (volume, variety, velocity) in satellite images has been defined by Mingmin Chi et al. [3], where they have used the 3V's dimensions of satellite image data in Trinity framework, this framework is proposed for a better understanding of data in terms of satellite images applications [3]. The 5V's dimensions of big data (volume, variety, velocity, value, and veracity) are defined by D S Singh and G Singh [7]:

- In terms of volume, the collection of satellite images from different sources (NASA, USGS, NRSC), and stored data are characterized by volume.
- In terms of variety, satellite images are generated from various sources like radar, optical and thermal sensors with multi-temporal (collected on different dates), multi-resolution (different spatial resolution), and multispectral (different bands data) satellite images.
- Velocity define the growth rate of the satellite images.
- In terms of value, the ability to analyze the data and provide a better understanding of information including various key areas such as the behavior and quality of satellite images.
- In terms of veracity, it refers to the accuracy of the truth data, uncertainly in the data can be caused for various reasons which may raise legal questions, privacy issues, duplication, etc. [7].

**Big Data Analytical Technique:** Big data analytics is a modern approach that can handle large and complex datasets, it can be useful to extract meaningful information from satellite images and it can apply to real-time data as well as offline datasets. There are some widely used big data analytical techniques are defined as machine learning, ensemble analysis, association rule learning, deep learning, precision analysis, and divide-and-conquer [5] which may beneficial while working with satellite images.

In this study, a machine learning algorithm (KNN) is used as a similarity measurement algorithm.

The similarity measurement algorithm is used to find the similarity between two satellite images using satellite image features, and based on the similarity result system predicts that images are similar or not. The KNN algorithm calculates the Euclidean distance between retrieved image features and features that is store in the feature database.

Euclidean distance between two variables X and Y is defined as [31]

$$\text{Distance (x,y)} = (\sum |X_i - Y_i|^2)^{1/2} \quad (8)$$

The Euclidean distance metric is most commonly used as similarity measurement algorithm in image retrieval system because of its efficiency and effectiveness. It measures the distance between two vectors of images by calculating the square root of the sum of the squared absolute differences [31].

### Model Development

Many researchers have developed various methods, frameworks, and platforms for satellite image retrieval. The traditional satellite image retrieval system does not indicate the quality of the satellite images before downloading the particular image. Hence, user can decide that image is suitable or not for post-processing after retrieving the image. The image retrieval system aims to retrieve images from the image repository as per the user requirements. Every image retrieval system has its image repository; the image repository contains a variety of images. In the case of satellite image repository, some images may be affected due to weather conditions or cloud covers, therefore, these types of images are not suitable for post-processing because they cannot produce an efficient result for interpretation or analysis of earth surface monitoring.

In this present study, a feature-based image retrieval (FBIR) system has been developed which can indicate the quality of the satellite image before the retrieval of an image. The developed system minimizes the downloading time, wastage of the internet, and the most important thing is the time of the user that is consumed during the pre-processing of raw images. The FBIR System is an intelligent image retrieval system which uses image features result for satellite image retrieval from the image repository that can produce an



efficient result for interpretation and analysis of earth surface monitoring. The flow chart of developed FBIR is shown in figure 3.

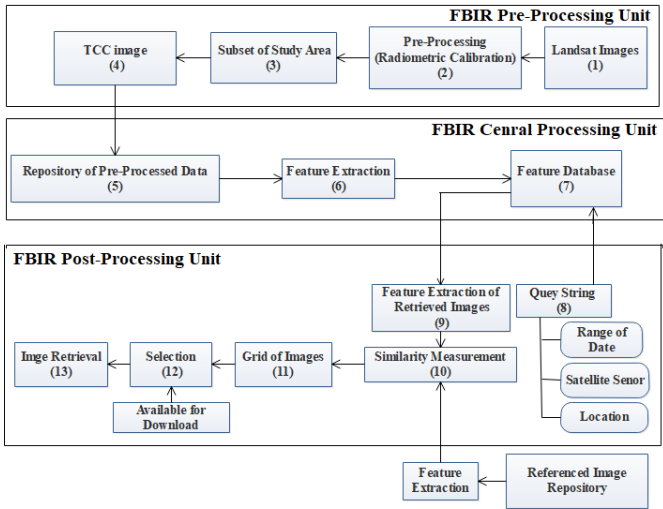


Figure 3. Feature-Based Image Retrieval System.

FBIR Pre-Processing Unit

FBIR pre-processing unit is an upper layer and an important phase of the system, the objective of this unit is to collect satellite images from different sources and prepare a feature database using satellite image features values for further use of FBIR central processing unit. FBIR pre-processing unit is a combination of several subunits Landsat images, Pre-processing, Subset of study area, and TCC images.

FBIR Central Processing Unit

FBIR central processing unit is the middle layer of the system; the FBIR central processing unit builds a virtual relationship between FBIR pre-processing unit and the FBIR post-processing unit for better image retrieval. FBIR central processing unit is a combination of several subunits such as

Repository of pre-processed data, Feature extraction, and Feature database.

FBIR Post-Processing Unit

The FBIR post-processing unit is the bottom layer of this system, the FBIR post-processing unit is responsible for the input and output of the system. Query string and image retrieval are the subunits of the FBIR post-processing unit and act as an input and output, and the other subunits are feature extraction of retrieved images, similarity measurement, grid of images, and selection.

In proposed system the main challenging task is to identify the better image for download. For this purpose a referenced image repository has been created in which features of some 2 best possible images of every year are stored. The referenced images are the best possible image of every year which can produce a better result for post-processing, and the selection of referenced images depends on the classification result. To identify 2 best possible images of every year, ML, ANN, and SVM classification techniques has been applied on various vegetation indices. To perform the classification vegetation indices such as NDVI, SAVI, AWEI and NDBI has been calculated as their mathematical formulation is given in equations 4 to 8 by using the Red, NIR, GREEN, SWIR1 and SWIR2 bands. Further, indices images have been layer staked in ENVI 5.1 and then ML, ANN and SVM classification techniques were applied. After performing the classification on all the 140 images, total 10 images from 2016 to 2020 (2 best possible images of each year) have been found by carrying out the accuracy measurement. Overall accuracy and Kappa coefficient of best images attained through ML, ANN and SVM classifiers are given in Table 2.

Table 2. Details of Referenced Images

Image ID	Maximum Likelihood		Artificial Neural Network		Support Vector Machine	
	Overall accuracy	Kappa Coefficient	Overall accuracy	Kappa Coefficient	Overall accuracy	Kappa Coefficient
IMG_1	80.81%	0.7369	80.28%	0.7235	83.63%	0.7648
IMG_2	80.28%	0.7269	84.15%	0.7788	85.74%	0.7958
IMG_3	80.81%	0.7369	80.28%	0.7235	83.63%	0.7648
IMG_4	82.22%	0.7538	83.98%	0.7746	85.04%	0.7845
IMG_5	79.75%	0.7203	82.22%	0.7499	82.57%	0.7516
IMG_6	83.39%	0.7702	84.59%	0.7851	86.64%	0.8115
IMG_7	84.93%	0.7915	83.73%	0.7753	84.59%	0.7833

IMG_8	80.28%	79.23%	79.23%	0.7133	82.39%	0.7486
IMG_9	83.10%	0.763	78.35%	0.7027	85.21%	0.7889
IMG_10	80.63%	0.734	84.33%	0.7789	85.92%	0.7994
IMG_11	82.3944	0.7569	80.99%	0.7293	85.21%	0.7884
IMG_12	83.63%	0.7743	82.92%	0.7613	86.62%	0.8086
IMG_13	83.98%	0.7772	82.57%	0.7562	85.92%	0.7995
IMG_14	82.39%	0.7544	85.56%	0.7925	85.39%	0.7905

From Table 2, it is observed that overall accuracy and kappa coefficient attained through SVM classifier for all the classified images are higher than the ML and ANN classifiers. Therefore, it is found that SVM classifier is producing the better classified images to create a referenced images repository. Here, in this paper, SVM classifier results are considered to assess the best available images for retrieval and download via FBIR system.

### Implementation of Feature Bases Image Retrieval System

#### Collection of Satellite Images

In this current study, satellite images were collected from USGC earth explorer web portal for building an offline image repository. Approximately 140 Landsat-8 OLI images have been stored in the image repository to develop a FBIR system.

#### Pre-Processing of Satellite Images

Satellite image pre-processing includes the radiometric calibration, which is used to convert the digital numbers of an image into reflectance values.

#### Subset of Study Area

After the pre-processing of all the images, there is need to extract the study area from all the images. The study areas are located between latitude 26.945746° to 26.734630° and longitude between 80.74451° to 81.142585°.

#### TCC Image

The true color composite (TCC) image has been generated from pre-processed images. The red, green, and blue bands have been used to generate the true color composite image which shows the shapes and sizes of the various features.

#### Pre-Processed Data Repository

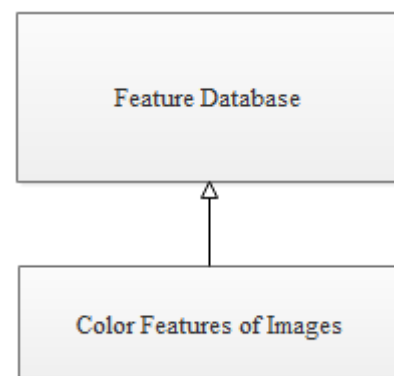
The pre-processed data repository contains the true color composite images that have been pre-processed for color features extraction.

#### Feature Extraction

The image feature is playing an important role in FBIR system. The color moment feature extraction technique is used to extract color features from the true color composite image and it has extracted features are stored in the database to build a feature database.

#### Feature Database

The process of building the feature database depends on the image features results which are extracted by the color moment algorithm from the image. The feature values need to be stored in the feature database for further processing as shown in Figure 4.



**Figure 4. Feature Database**

Many image feature extraction techniques are available, but in this study, color features (i.e. Red, Green and Blue) have been considered to develop FBIR System. The color moment algorithm is used to extract color features from the image. After feature extraction, the extracted feature values were stored in the feature database.

Query String

The query string acts as an input of this system and it is a combination of a range of date, name of satellite sensor, and location of the study region as shown in Figure 5.

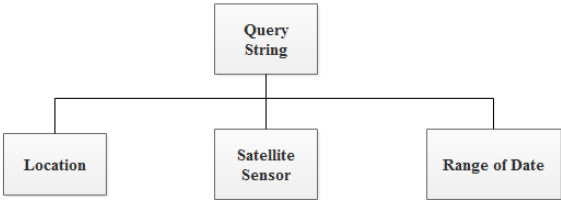


Figure 5. Query String.

The range of date contains the start date and end date of image to be retrieved, the satellite sensor contains the name of the satellite, and the location contains the name of the place of the study region.

Feature Extraction of Retrieved Image

The query string interacts with the feature database, and the feature database returns the feature values of images based on the query string. A similarity measurement algorithm has been applied on the obtained feature values of images in the query string with the feature values retrieved from the referenced image repository.

Similarity Measurement

The referenced image repository has two best images of every year which helps to find better images from features values of images that have been returned by the feature database. The system matches the feature values of images with the features of the best image of that year (referenced image), and based on similarity result, the system predicts that the image is good for post-processing or not. In this study, the KNN algorithm is used as a similarity measurement algorithms. The KNN matches the feature values of images that have been returned by the feature database with the features of the best possible image of that year.

Grid of Images

The images have been filtered after the feature matching on the basis of similarity result, and listed in a grid as shown in Figure 6.

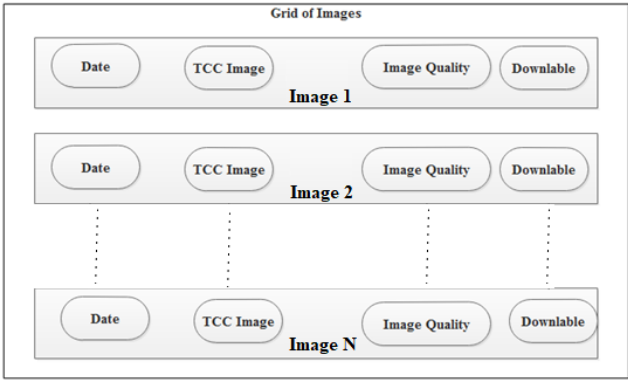


Figure 6. Grid of Images.

For image retrieval, the user gives a query string as an input. Query string interacts with the feature database, and it returns feature values based on a query string. The similarity measurement algorithm is used to find better images from it, the similarity measurement algorithm matches the features of a referenced image with the feature values of images which have returned by the feature database for finding similarity, and based on the similarity result the system predicts that image is better for post-processing or not. After the similarity measurement, the system returns a grid of images based on query string. The user can select an image from grid for retrieval, and the system only allows to select those image from image grid which are better for post-processing and can produce efficient result for earth surface monitoring.

Selection

The user can select an image from the grid of images for retrieval which one user want to download.

Image Retrieval

The system serves the image which is selected by a user in the selection phase for post-processing as a result.

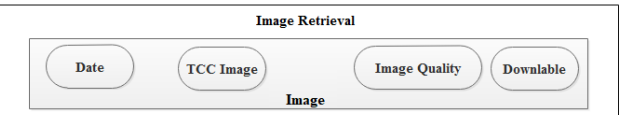


Figure 7. Image Retrieval.

In image retrieval module, end user will be able to download the suggested images and further can use download images as per their requirements.



Result and Discussion

The main objective of this work is to develop such type of feature based image retrieval system (FBIR), which can suggest the best possible available images for end users to directly use downloaded images as per their requirements. In order to develop a FBIR system, two Landsat-8 OLI image of 27 January 2020 (Image ID: IMG\_12) and 12 February 2020 (Image ID: IMG\_13) as given Table I have been used.

In the initial phase of the FBIR development, Landsat-8 OLI satellite images were collected to build an offline image repository. FBIR image repository contains approximately 140 satellite images from February 17, 2016 to April 16, 2020, which were pre-processed one by one and stored in the Pre-processed image repository. In the pre-processing phase, radiometric calibration has been applied on each satellite image which is stored in the image repository to convert the digital number of an image into surface reflectance values. After pre-processing on all the images, area of interest has been cropped from these images and then true color composite (TCC) images have been generated one by one using the Red, Green, and Blue bands and generated TCC images are stored in the Pre-processed Image Repository. TCC images generated through IMG\_12 and IMG\_13 are shown in Figure 8 (a) and 8 (b):

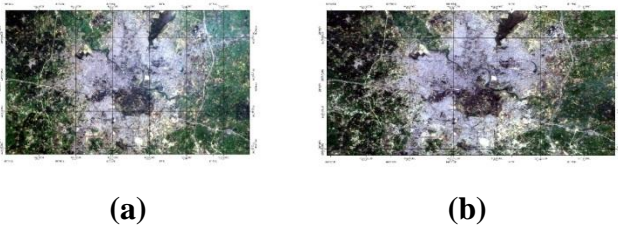


Figure 8. TCC Image of Study Area (a) Image ID: IMG\_12 and (b) Image ID: IMG\_13.

True-color composite (TCC) image is a combination of the red, green, and blue bands. The TCC image defines what would be observed naturally by the human eye such as the vegetation appears green, the water appears blue to black and bare ground and the impermeable surface appears light gray and brown.

In the second phase of FBIR system development, color moment algorithm has been applied for color feature extraction on each TCC image and retrieved color feature values are stored in a

feature database under FBIR central processing unit. Color histogram of Red, Green, and Blue bands obtained from IMG\_12 and IMG\_13 images are shown in Figure 9 (a) and b (b):

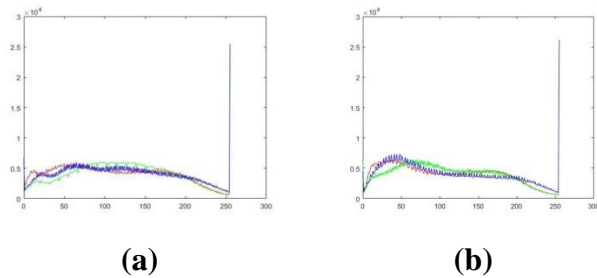


Figure 9. Color Histogram of TCC image (a) Image ID: IMG\_12 and (b) Image ID: IMG\_13.

Feature values are retrieved from IMG\_12, and IMG\_13 as MeanR, MeanG, MeanB, StdR, StdG and StdB of Red, Green and Blue bands respectively are mentioned in Table 3.

Table 3. Color Features of Sample Images (Development Phase)

Image ID	Mean R	Mean G	Mean B	StdR	StdG	StdB
IMG_12	113	122.84	120.41	67.28	62.00	67.60
IMG_13	108.79	112.55	111.45	68.30	65.01	70.36

Similarly, color moment feature extraction technique has been applied on all 140 images and the retrieved feature values like MeanR, MeanG, MeanB, StdR, StdG and StdB of Red, Green and Blue bands respectively are stored in the feature database as shown in Figure 10.

Data	Acquisition_ID	Acquisition_Date	Tcc	Location	State	Country	top_left_lat	top_left_lon	bottom_right_lat	bottom_right_lon	MeanR	MeanG	MeanB	stdR	stdG	stdB
Landsat-8	LC08_L1TP_144041_20191023_20191030_01_T1	2019-10-23	tcc/23oct2019	Lucknow	Uttar Pradesh	India	26.945748?	80.744511?	26.734630?	81.142585?	78.969	78.199	78.316	65.501	65.444	65.671
Landsat-8	LC08_L1TP_144041_20191108_20191115_01_T1	2019-11-08	tcc/8nov2019	Lucknow	Uttar Pradesh	India	26.945748?	80.744511?	26.734630?	81.142585?	115.75	109.54	100.2	56.479	51.088	49.667
Landsat-8	LC08_L1TP_144041_20191124_20191203_01_T1	2019-11-24	tcc/24nov2019	Lucknow	Uttar Pradesh	India	26.945748?	80.744511?	26.734630?	81.142585?	147.51	149.14	148.13	58.915	56.975	57.243
Landsat-8	LC08_L1TP_144041_20191210_20191217_01_T1	2019-12-10	tcc/10dec2019	Lucknow	Uttar Pradesh	India	26.945748?	80.744511?	26.734630?	81.142585?	142.42	146.5	141.6	61.835	60.222	61.313
Landsat-8	LC08_L1GT_144041_20191229_20200110_01_T2	2019-12-26	tcc/26dec2019	Lucknow	Uttar Pradesh	India	26.945748?	80.744511?	26.734630?	81.142585?	149.24	149.97	152.26	55.727	55.626	55.626
Landsat-8	LC08_L1TP_144041_20200111_20200114_01_T1	2020-01-11	tcc/11jan2020	Lucknow	Uttar Pradesh	India	26.945748?	80.744511?	26.734630?	81.142585?	158.55	158.8	161.67	53.885	53.942	54.371
Landsat-8	LC08_L1TP_144041_20200127_20200210_01_T1	2020-01-27	tcc/27jan2020	Lucknow	Uttar Pradesh	India	26.945748?	80.744511?	26.734630?	81.142585?	113	122.84	120.41	67.287	62.006	67.606
Landsat-8	LC08_L1TP_144041_20200212_20200225_01_T1	2020-02-12	tcc/12feb2020	Lucknow	Uttar Pradesh	India	26.945748?	80.744511?	26.734630?	81.142585?	108.79	112.55	111.45	68.308	65.01	70.365
Landsat-8	LC08_L1TP_144041_20200315_20200325_01_T1	2020-02-28	tcc/28feb2020	Lucknow	Uttar Pradesh	India	26.945748?	80.744511?	26.734630?	81.142585?	75.93	74.159	72.956	69.888	70.382	70.72
Landsat-8	LC08_L1TP_144041_20200328_20200331_01_T1	2020-03-15	tcc/15march2020	Lucknow	Uttar Pradesh	India	26.945748?	80.744511?	26.734630?	81.142585?	82.53	76.626	67.991	56.144	49.698	49.975
Landsat-8	LC08_L1TP_144041_20200331_20200410_01_T1	2020-03-31	tcc/31march2020	Lucknow	Uttar Pradesh	India	26.945748?	80.744511?	26.734630?	81.142585?	108.55	103.92	102.25	64.575	66.467	66.768
Landsat-8	LC08_L1TP_144041_20200416_20200416_01_RT	2020-04-16	tcc/16april2020	Lucknow	Uttar Pradesh	India	26.945748?	80.744511?	26.734630?	81.142585?	132.73	122.24	124.09	63.657	63.867	63.016
Landsat-8	LC08_L1TP_144041_20181105_20181115_01_T1	2018-11-05	tcc/5Nov2018	Lucknow	Uttar Pradesh	India	26.945748?	80.744511?	26.734630?	81.142585?	130.7	127.76	124.19	63.646	63.488	64.649
Landsat-8	LC08_L1GT_144041_20180105_20180118_01_T2	2018-01-05	tcc/5Jan2018	Lucknow	Uttar Pradesh	India	26.945748?	80.744511?	26.734630?	81.142585?	138.69	139.93	141.56	57.683	57.714	57.774
Landsat-8	LC08_L1TP_144041_20181207_20181211_01_T1	2018-12-07	tcc/7Dec2018	Lucknow	Uttar Pradesh	India	26.945748?	80.744511?	26.734630?	81.142585?	129.37	134.09	137.97	60.983	60.26	61.043
Landsat-8	LC08_L1TP_144041_20180411_20180417_01_T1	2018-04-11	tcc/11Apr2018	Lucknow	Uttar Pradesh	India	26.945748?	80.744511?	26.734630?	81.142585?	166.35	165.88	166.6	50.888	50.894	51.379
Landsat-8	LC08_L1TP_144041_20180513_20180517_01_T1	2018-05-13	tcc/13May2018	Lucknow	Uttar Pradesh	India	26.945748?	80.744511?	26.734630?	81.142585?	137.53	138.23	144.25	63.531	62.409	63.705
Landsat-8	LC08_L1TP_144041_20180614_20180703_01_T1	2018-06-14	tcc/14June2018	Lucknow	Uttar Pradesh	India	26.945748?	80.744511?	26.734630?	81.142585?	153.85	152.41	153.6	54.324	55.226	57.213
Landsat-8	LC08_L1TP_144041_20180121_20180206_01_T1	2018-01-21	tcc/21Jan2018	Lucknow	Uttar Pradesh	India	26.945748?	80.744511?	26.734630?	81.142585?	115.5	122.24	131.43	67.72	64.092	67.794
Landsat-8	LC08_L1TP_144041_20181121_20181121_01_RT	2018-11-21	tcc/21Nov2018	Lucknow	Uttar Pradesh	India	26.945748?	80.744511?	26.734630?	81.142585?	134.83	137.05	143.77	61.272	60.712	60.731
Landsat-8	LC08_L1TP_144041_20180222_20180308_01_T1	2018-02-22	tcc/22Feb2018	Lucknow	Uttar Pradesh	India	26.945748?	80.744511?	26.734630?	81.142585?	112.72	115.66	113.69	68.868	66.027	68.693
Landsat-8	LC08_L1TP_144041_20180427_20180502_01_T1	2018-04-27	tcc/27Apr2018	Lucknow	Uttar Pradesh	India	26.945748?	80.744511?	26.734630?	81.142585?	137.07	138.63	139.22	63.088	63.088	64.292
Landsat-8	LC08_L1TP_144041_20180529_20180605_01_T1	2018-05-29	tcc/29May2018	Lucknow	Uttar Pradesh	India	26.945748?	80.744511?	26.734630?	81.142585?	135.01	135.18	141.24	64.283	62.625	64.669

Figure 10. Color Features of Images.

To retrieve the best available image from FBIR system, graphical user interface (GUI) as a web portal has been developed. The main objective behind the development of GUI is that end user can easily find the best available images of the required date; if that particular date image is not recommended to use or not available than FBIR system will also suggest the best available images of its near by dates. The developed GUI for query string is shown in Figure 11, in which user gives a query string as input in the system. A query string is a combination of a range of acquisition date (start and end date), name of satellite sensor, and location of required image.

LOCATION	<input type="text" value="Lucknow"/>
SATELLITE	<input type="text" value="Landsat-8"/>
DATE	<input type="text" value="01/01/2020"/> <input type="text" value="04/16/2020"/>
<input type="button" value="SEARCH"/>	

Figure 11. GUI of a Query String (Development Phase).

FBIR system receives a query string as an input and forwards it to the feature database, the feature database returns the feature values of images according to the query string. The FBIR system also has a referenced image repository that contains the two quality images of every year. The referenced image helps to find the better image from the feature values of images that are returned by the feature database. The similarity measurement algorithm matches the feature values

of images with referenced image features and based on the similarity result the system predicts that the image is better or not.

In this study, KNN has used as a similarity measurement algorithm in which Euclidian distance is used to calculate the distance between two images. Euclidian distance measures the distance between two vectors of images by calculating the square root of the sum of the squared absolute differences [31].

In order to test the proposed FBIR system, initially two Landsat-8 OLI images (i.e. Img\_12 and Img\_13) have been taken. The FBIR system will extract the features of Img\_12 and Img\_13 images, and will compute the difference between two features vectors as mentioned in table 4.

Table 4. Similarity Measurement Table (Development Phase)

Band	IMG_12	IMG_13	Result
	X	Y	R=(X-Y) <sup>2</sup>
Red Mean	113	108.79	17.72
Green Mean	122.84	112.55	105.88
Blue Mean	120.41	111.45	80.28
Red Std.	67.287	68.308	1.042
Green Std.	62.006	65.01	9.024
Blue Std.	67.606	70.365	7.612

From Table 4, X variable contains the color feature values of Img\_12 image, while Y variable contains the color feature values of Img\_13

image, R contains the squared absolute differences (X-Y) 2.

In this study, the square root of sum of distance between features vectors of two different images has been considered to compute the similarity score as given in equation (9), which is used to measure the similarity between two images. If the similarity score between two image feature vectors is 0, it means images are same on the basis of their features. The detail of similarity score's interpretation for various retrieved scores are given in Table 5.

**Table 5. Property of interpretation of similarity score**

Similarity Score	Properly
0-10	Perfectly similar, and best for post-processing.
11-15	Partial similar, and good for post-processing.
16-20	Partial similar, and okay for post-processing.
>20	Not similar

Similarity score contains the square root of the sum of the squared absolute differences (R), the formula of similarity measurement is given below:

**Similarity Score**  
$$= \sqrt{R_1 + R_2 + R_3 + R_4 + R_5 + R_6} \quad (9)$$

Where, R\_1 is the squared absolute differences of Read Mean features, R\_2 is the squared absolute differences of Green Mean features, R\_3 is the squared absolute differences of Blue Mean features, R\_4 is the squared absolute differences of Read Standard Deviation features, R\_5 is the squared absolute differences of Green Standard Deviation features, and R\_6 is the squared absolute differences of Blue Standard deviation features.



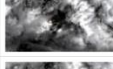


From table 5 and eqn(8), the similarity score between IMG\_12 and IMG\_13 is given below:

**S1**  
$$= \sqrt{17.72 + 105.88 + 80.28 + 1.042 + 9.024 + 7.612}$$
$$= 14$$

From Table 4, eqn(9), it observed that the similarity score between 27 January 2020 (Image ID: IMG\_12) and 12 February 2020 (Image ID:

IMG\_13) images is between 11 to 15 hence can say both images are partial similar, and good for post-processing.

After the analysis of similarity score, FBIR system will list all the best available images in the image grid and the system will suggest to download only those images whose similarity score lies between 0 to 20 as shown in Figure 12.

DATE	SAMPLE IMAGE	IMAGE QUALITY	DOWNLOAD
2020-04-16		Similarity Score: 20 partial Similar, and okay for post-processing.	DOWNLOAD
2020-03-31		Similarity Score: 27 not similar, hence we suggest to select those image which is similar to a reference image.	
2020-03-15		Similarity Score: 80 not similar, hence we suggest to select those image which is similar to a reference image.	
2020-02-28		Similarity Score: 77 not similar, hence we suggest to select those image which is similar to a reference image.	
2020-02-12		Similarity Score: 14 Similar, and good for post-processing.	DOWNLOAD

**Figure 12. GUI-Grid of Images (Development Phase).**

After clicking on the download button, the portal will be redirected to the final download page and will ask for the final confirmation for the download.

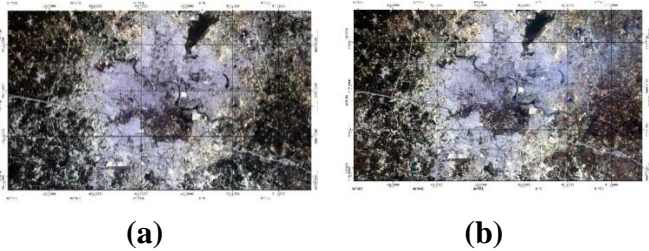


**Figure 13. GUI- Image Retrieval.**

Finally, the web portal will ask for the final conformation before downloading the image.

**Testing of FBIR System**

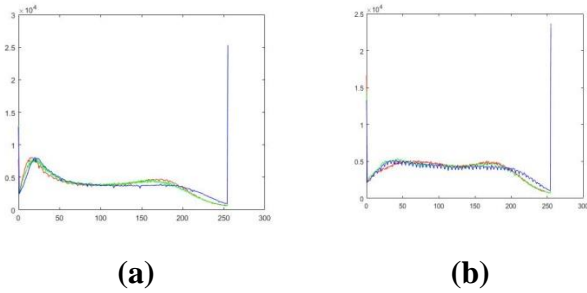
Two Landsat-8 OLI images of 13 March 2019 (Image ID: IMG\_9) and 29 March 2019 (Image ID: IMG\_10) as given Table I, have been used for the testing of the FBIR System. TCC images generated through IMG\_9 and IMG\_10 are shown in Figure 14 (a) and 14 (b):



**Figure 14. TCC Image of Study Area (a) Image ID: IMG\_9 (b) Image ID: IMG\_10.**



Color histogram of Red, Green, and Blue bands obtained from IMG\_9 and IMG\_10 images are shown in Figure 15 (a) and 15 (b), and feature values are retrieved from IMG\_9 and IMG\_10 as MeanR, MeanG, MeanB, StdR, StdG and StdB of Red, Green and Blue bands respectively are mentioned in Table 6.



**Figure 15. Color Histogram of TCC image (a) Image ID: IMG\_9 and (b) Image ID: IMG\_10.**

**Table 6. Color Features of Sample Images (Testing Phase)**

Image ID	MeanR	MeanG	MeanB	StdR	StdG	StdB
IMG_9	107.33	105.07	108.58	71.158	70.61	73.369
IMG_10	116.65	114.81	119.1	67.733	68.577	70.462

A query string GUI is shown in Figure 16, in which an input range of dates has been given to test the performance of proposed FBIR system.

LOCATION	<input type="text" value="Lucknow"/>
SATELLITE	<input type="text" value="Landsat-8"/>
DATE	<input type="text" value="01/01/2019"/> <input type="text" value="12/31/2019"/>
<input type="button" value="SEARCH"/>	

**Figure 16. GUI of a Query String (Testing Phase)**

The referenced image of input year (2019) helps to find the better image from the feature values of images that are returned by the feature database using similarity measurement algorithm and the system predicts that the image is better or not using similarity score as given in Table 5.

**Table 7. Similarity Measurement Table (Testing Phase)**

Band	IMG_9 X	IMG_10 Y	Result R=(X-Y) <sup>2</sup>
Red Mean	107.33	116.65	86.86
Green Mean	105.07	114.81	94.86
Blue Mean	108.58	119.1	110.67
Red Std.	71.158	67.733	11.73
Green Std.	70.61	68.577	4.13
Blue Std.	73.369	70.462	8.43

From Table 7, X variable contains the color feature values of IMG\_9 image, while Y variable contains the color feature values of IMG\_10 image, R contains the squared absolute differences (X-Y) <sup>2</sup>.

By using the value of R mentioned in table 7 and eqn (8), the similarity score between images IMG\_9 and IMG\_10 is calculated as follows:

$$S2 = \sqrt{86.86 + 94.86 + 110.67 + 11.73 + 4.13 + 8.43} = 17$$

the similarity score between both the images is between 16 to 20. Hence, it is observed that both the images are partially similar, and okay for post-processing. after the analysis of similarity score, FBIR system will list all the best available images in the image grid and the system will suggest to download only those images whose similarity score lies between 0 to 20 as shown in Figure 17.

DATE	SAMPLE IMAGE	IMAGE QUALITY	DOWNLOAD
2019-04-14		Similarity Score: 34 not similar, hence we suggest to select those image which is similar to a reference image.	
2019-03-29		Similarity Score: 17 partial Similar, and okay for post-processing	<input type="button" value="DOWNLOAD"/>
2019-03-13		Similarity Score: 0 Perfectly Similar, and better for post-processing.	<input type="button" value="DOWNLOAD"/>
2019-02-09		Similarity Score: 5 Perfectly Similar, and better for post-processing.	<input type="button" value="DOWNLOAD"/>
2019-01-24		Similarity Score: 123 not similar, hence we suggest to select those image which is similar to a reference image.	

**Figure 17. GUI-Grid of Images (Testing Phase).**

After clicking on the download button, the portal will be redirected to the final download page and will ask for the final confirmation for the download.



Figure 18. GUI- Image Retrieval.

Validation of the FBIR System

Two Landsat-8 OLI images of 21 January 2018 (Image ID: IMG\_6) and 23 December 2018 (Image ID: IMG\_7) as given Table I, have been used for the validation of the FBIR system. TCC images generated through IMG\_6 and IMG\_7 are shown in Figure 19 (a) and 19 (b):

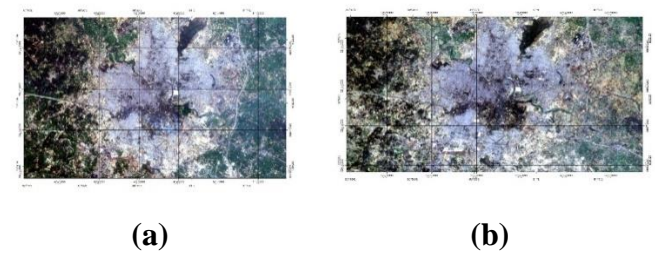


Figure 19 TCC Image of Study Area (a) Image ID: IMG\_6 and (b) Image ID: Img\_7.

Color histogram of Red, Green and Blue bands obtained from IMG\_6 and IMG\_7 images are shown in Figure 20 (a) and 20 (b), and feature values are retrieved from IMG\_6 and IMG\_7 as MeanR, MeanG, MeanB, StdR, StdG and StdB of Red, Green and Blue bands respectively are mentioned in Table 8.

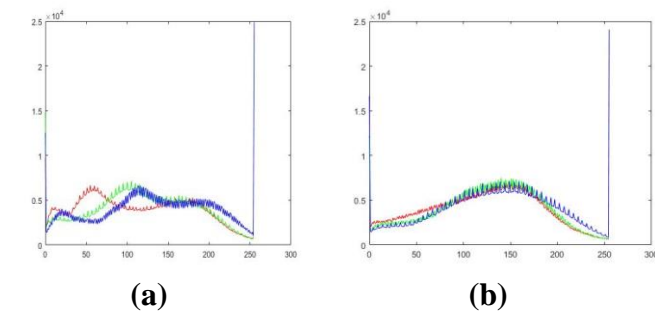


Figure 20 Color Histogram of TCC image (a) Image ID: IMG\_6 and (b) Image ID: IMG\_7.

Table 8. Color Features of Sample Images (Validation Phase)

Image ID	Mean R	Mean G	Mean B	StdR	StdG	StdB
IMG_6	115.5	122.24	131.43	67.72	64.09	67.79
IMG_7	122.32	127.3	132.49	62.03	60.34	62.92

The developed GUI is shown in Figure 21, in which user gives a query string for validation of the system as input.

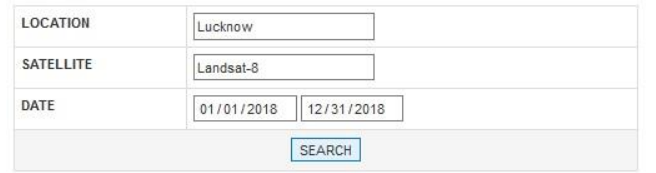


Figure 21 GUI of a Query String (Validation Phase).

The referenced image of that year (2018) helps to find the better image from the feature values of images that are returned by the feature database using similarity measurement algorithm and system predicts that the image is better or not using similarity score.

Table 9. Similarity Measurement Table (Validation Phase).

Band	IMG_6 X	IMG_7 Y	Result R=(X-Y) <sup>2</sup>
Red Mean	115.5	122.32	46.51
Green Mean	122.24	127.3	25.60
Blue Mean	131.43	132.49	1.12
Red Std.	67.72	62.031	32.36
Green Std.	64.092	60.347	14.02
Blue Std.	67.794	62.922	23.73






From Table 9, X variable contains the color feature values of IMG\_6 image, while Y variable contains the color feature values of IMG\_7 image, R contains the squared absolute differences (X-Y)<sup>2</sup>.

From table 9 and eqn (8), the similarity score between images IMG\_6 and IMG\_7 is given bellow:

$$S3 = \sqrt{46.51 + 25.60 + 1.12 + 32.36 + 14.02 + 23.73} = 11$$
 (11)

the similarity score between 21 January 2018 (Image ID: IMG\_6) and 23 December 2020 (Image ID: IMG\_7) images is between 11 to 15 hence can say both images are partial similar, and good for post-processing. After the analysis of similarity score, FBIR system will list all the best available images in the image grid and the system will suggest to download only those images whose

similarity score lies between 0 to 20 as shown in Figure 22.

DATE	SAMPLE IMAGE	IMAGE QUALITY	DOWNLOAD
2018-12-23		Similarity Score: 11 Similar, and good for post-processing.	DOWNLOAD
2018-12-07		Similarity Score: 21 not similar, hence we suggest to select those image which is similar to a reference image.	
2018-11-21		Similarity Score: 29 not similar, hence we suggest to select those image which is similar to a reference image.	
2018-11-05		Similarity Score: 18 partial Similar, and okay for post-processing.	DOWNLOAD
2018-10-20		Similarity Score: 242 not similar, hence we suggest to select those image which is similar to a reference image.	

**Figure 22 GUI-Grid of Images (Validation Phase).**

After clicking on the download button, the portal will be redirected to the final download page and will ask for the final confirmation for the download.

Date: 2018-12-23

Please Click on Download Button for Image Retrieval

DOWNLOAD

**Figure 23 GUI- Image Retrieval (Validation Phase).**

In this paper, feature based image retrieval (FBIR) system has been proposed to assess the satellite image quality using big data analytical techniques. The developed system minimizes the downloading time, wastage of the internet, and the most important thing is the time of the user that is consumed during the pre-processing of raw images. The FBIR System is an intelligent image retrieval system which uses image features result for satellite image retrieval from the image repository that can produce an efficient result for interpretation and analysis of earth surface monitoring.

### Conclusion

The satellite image carries the characteristics of big data as we know that satellite data are increasing all over the world day by day. Currently, data is increasing in high volume and named as big data, and big data analysis makes it easy to analyze large and complex datasets. Big data analysis is also used in decision making, mining, and predictive analysis. Traditional data analysis methods are not designed for large volume and complex datasets, hence these data

analysis methods are less capable to compute large datasets. Big Data analysis is a modern approach that overcomes the challenges of traditional analysis approaches. The optical satellite images undertaken in this paper maybe weather affected and cloud affected. Hence, affected images may not be suitable for post-processing because they cannot produce an efficient result for interpretation and analysis of earth surface monitoring. In this present study, FBIR system has been developed by considering the Lucknow as study region, which is useful to find out the best available satellite image from the image repository for post-processing and especially does not serve the affected images. In future, FBIR system can be extended at District, State and Country level, and FBIR may become very beneficial for the end users to retrieve less affected satellite images from the image repository, which could be more suitable for further processing.

### Acknowledgement

Authors are thankful to the Advanced Computing & Research Laboratory, Department of Computer Application, Integral University Lucknow, India for providing the support to this work.

### References

- [1] Ahmed, T., Singh, D., & Raman, B. (2016). Potential application of Kanade–Lucas–Tomasi tracker on satellite images for automatic change detection. *Journal of Applied Remote Sensing*, 10(2), 026018-1-026018-18.
- [2] Ma, Y., Wu, H., Wang, L., Huang, B., Ranjan, R., Zomaya, A & Wei. (2015). Remote sensing big data computing: Challenges and opportunities. *Future Generation Computer Systems*, 51(1), 47-60.
- [3] Chi, M., Plaza, A., Benediktsson, J.A., Sun, Z., Shen, J. & Zhu, Y. (2016). Big Data for Remote Sensing: Challenges and Opportunities. *Proceedings of the IEEE*. 104(11), 2207-2219.
- [4] Ahuja, K. & Jani, N.N. (2016). A Study Of Traditional Data Analysis And Sensor Data Analytics. *International Journal of Information Sciences and Techniques (IJIST)*. 6(2), 185-190.



- [5] Shu, H. (2016). Big data analytics: six techniques. *Geo-spatial Information Science*. 119-128.
- [6] Pandya, R., Sawant, V., Mendjoge, N. & D'silva, M. (2015). Big Data Vs Traditional Data. *International Journal for Research in Applied Science & Engineering Technology*. 3(x), 192-196.
- [7] Singh, D.S., & Singh, G. (2017). Big data – A Review. *International Research Journal of Engineering and Technology (IRJET)*. 4(4), 822-824.
- [8] Guo, H. (2017), Big Earth data: A new frontier in Earth and information sciences. *Big Earth Data*. 4-20.
- [9] Tawade, M.S.S. (2018). Applications of Big Data: Review Paper. *International Research Journal of Engineering and Technology*. 5(2), 1631-1633.
- [10] Gunturi, Y.K, & Raju, K.K. (2017). RealBDA: A Real Time Big Data Analytics For Remote Sensing Data By Using Mapreduce Paradigm. *International Journal Of Engineering Sciences & Research Technology*. 6(1), 435-442.
- [11] Rathore, M.M.U., Paul, A., Ahmad, A., Chen, B.W., Huang, B. & Ji, W. (2015). Real-Time Big Data Analytical Architecture for Remote Sensing Application. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 8(10), 4610 - 4621.
- [12] Patil, S., Patil, D.R. (2016). Real-Time Big Data Analytical Architecture For Image Processing Framework. *International Journal of Advanced Technology in Engineering and Science*. 4(6), 374-380.
- [13] Wulder, M.A., White, J.C., Loveland, T.R., Woodcock, C.E., Belward, A.S., Cohen, W.B., Fosnight, E.A., Shaw, J., Masek, G. & Roy, D.P. (2016). The global Landsat archive: Status, consolidation, and direction. *Remote Sensing of Environment*. 185, 271-283.
- [14] Kumar, R., & Ancy S. (2014). A Series of Earth Resource Satellites and its Potentialities. *International Journal of Advancements in Research & Technology*. 3(4), 4-9.
- [15] Montanaro, M. (2014). Lunsford, A., Tesfaye, Z., Wenny, B., & Reuter, D. (2014). Radiometric Calibration Methodology of the Landsat 8 Thermal Infrared Sensor. *Remote Sens*. 6, 8803-8821.
- [16] Thome, K.J. (2001). Absolute radiometric calibration of Landsat 7 ETM+ using the reflectance-based method. *Remote Sensing of Environment*. 78(1-2), 27-38.
- [17] Arivazhagan, S., Anbazhagan, A. (2017). ASTER Data Analyses for Lithological Discrimination of Sittampundi Anorthositic Complex, Southern India. *Geosciences Research*. 2(3), 196-209.
- [18] Demirkesena, A.C., Hazeltonb, N.W.J. & Sauderc, D.M. (2004). Automating Interpretation of Geological Structures from Landsat Tm Multi-Spectral Images and Dems. *ISPRS*.
- [19] Sutojo, T., Tirajani, P.S., Setiadi, D.R.I.M., Sari, C.A. & Rachmawanto, E.H. (2017, Nov 1-2). CBIR for classification of cow types using GLCM and color features extraction. 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), p. 182-187.
- [20] Kaur, M., & Dhingra, S. (2017, Feb 10-11). Comparative analysis of image classification techniques using statistical features in CBIR systems. 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), p-265-270.
- [21] Anusha, M.V., Reddy, M.V.U., & Ramashri, D.T. (2014). Content Based Image Retrieval Using Color Moments and Texture. *International Journal of Engineering Research & Technology (IJERT)*. 3(2), 2812-2815.
- [22] Moorthi, S.M., Misra, I. Kaur, R. Darji, N.P. & Ramakrishnan, R. (2011, Sept 22-24). Kernel based learning approach for satellite image classification using support vector machine. 2011 IEEE Recent Advances in Intelligent Computational Systems. P-107-110.
- [23] Han, P. Han, B. Lu, X. Cong, R. & Sun, D. (2019). Unsupervised classification for PolSAR images based on multi-level feature extraction. *International Journal of Remote Sensing*. 534-548.
- [24] Kamavisdar, P., Saluja, S. & Agrawal, S. (2013). A Survey on Image Classification Approaches and Techniques. *International*

- Journal of Advanced Research in Computer and Communication Engineering. 2(1), 1005-1009.
- [25] Srivastava, S. (2020). A survey on satellite image classification approaches. International Journal of Advance Research, Ideas and Innovations in Technology. 6(2), 480-483.
  - [26] Maa, L., Liuc, Y., Zhanga, X., Yed, Y., Yind, G ., Johnsonf, B.A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. ISPRS Journal of Photogrammetry and Remote Sensing. 152, 166-177.
  - [27] Bouhennache, R., Bouden, R., Ahmed, A.T. & Chaddad, A. (2018). A new spectral index for the extraction of built-up land features from Landsat 8 satellite imagery. Geocarto International. 1531-1551.
  - [28] HUETE, A.R. (1988). A Soil-Adjusted Vegetation Index (SAVI). Remote Sensing Of Environment, 25(3), 295-309.
  - [29] Feyisa, G.L., Meilby, H., Fensholt, R., Proud, S.R. (2014). Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery. Remote Sensing of Environment. 140, 23-35.
  - [30] Bhandaria, A.K., Kumara, A., & Singh, G.K. (2012). Feature Extraction using Normalized Difference Vegetation Index (NDVI): A Case Study of Jabalpur City. Procedia Technology. 6, 612-621.
  - [31] Malik, F. Baharudin, B. (2013). Analysis of distance metrics in content-based image retrieval using statistical quantized histogram texture features in the DCT domain. Journal of King Saud University – Computer and Information Sciences. 25, 207-218.
  - [32] Pardede, J., Sitohang, B., Akbar, S. & Khodra, M.L. (2017, Nov 1-2). Comparison of Similarity Measures in HSV Quantization for CBIR. 2017 International Conference on Data and Software Engineering (ICoDSE). p 1-6.