

Machine Learning for Satellite Images Classification using Spectral Signature for Features Extraction

Zainab H.Jarrallah^{1, a)*}

Prof. Dr. Maisa'a Abid Ali Khodher^{2, b)}

^{1,2} *Department of Computer Science, University of Technology, Baghdad, Iraq.*

^{a)*} zainabalahadly2021@gmail.com

^{b)} Maisa.A.Khodher@uotechnology.edu.iq

Abstract. When images are dedicated to identifying changes that have happened using techniques such as spectral signature, which may be used to extract features, they can be of great value. In this paper, propose using the spectral signature to extract information from satellite images then classify them into four classes. relied here on a set of data taken from the Kaggle website for satellite images representing different classes such as clouds, deserts, water, and green spaces. After pre-processing these images, transformed their data into a spectral signature using the Fast Fourier Transform algorithm (FFT), then reduced the data of each image by choosing the best 20 features and transformed it from a two-dimensional matrix to a one-dimensional vector using the Vector Quantization algorithm. All that after divided the data into training and testing, then entered it into four machine learning algorithms for classification (LR, RF, SGD, and NB) that classify satellite images, and the results were as follows: in logistic regression, the precision was 74%, the precision in random forest (RF) algorithm is 56%, in stochastic Gradient Descent (SGD) algorithm the precision is 81% and in Naive Bayes (NB) algorithm the precision is 57%.

Keywords. Image preprocessing, Fast Fourier Transform, Spectral signature, Machine learning, Data set

INTRODUCTION

At the local, regional, and global scales, Data taken from land cover images are important to facilitate the process of human interaction with nature. Image classification algorithms may be used to extract them from remotely sensed data[1]. Satellite remote sensing systems have created an archive of photographs of the planet that is becoming a more significant source of data for land cover and land-use change research[2].

The absorbance, reflectance, and transmittance of electromagnetic radiation are used to describe materials using spectral signatures or spectral fingerprints. These signatures are just graphs of an object's spectral reflectance as a function of wavelength[3]. introduce a novel technique in this Letter that enables the selective identification of land-cover changes by utilizing data from satellites[4] .

RELATED WORKS

In recent years, there were many studies focused on classification system based on spectral and machine learning algorithms:

In 2019 Ávila Vélez, et.al discovered the spectral signature of maize fields' ground coverings at various phases of growth (2 months, 2.3 months, and 4.3 months). In a similar vein, this study report recommends the use of a four-phase methodology: Georeferencing of maize crops, satellite picture selection, image radiometric calibration, and Maize crop spectral signature progress. At visible and near-infrared wavelengths, the spectral response or signature of maize fields was obtained, indicating considerable changes in crop growth. The utilization of satellite imaging becomes an intriguing instrument that presents an approach to more precise and regulated agricultural output monitoring systems[5] .

In 2020 Dyah R. et.al proposed When the amplitude of spectral variations between two observations due to land surface alteration is larger than any distortions, the spectral differences can be utilized to indicate a change. With an emphasis on multispectral pictures, the report analyzes achievements in bitemporal and multitemporal two-dimensional CD (change detection). It also goes through some of the CD methods utilized in synthetic aperture radar (SAR). The necessity of data selection and preparation for CD is a good place to start the conversation. The change analysis products that CD approaches might provide are then categorized to aid users in finding appropriate procedures for their applications. the evaluation reveals that critical and creative advancements in multispectral image analysis are being developed. For understanding the context of improvements, advantages, constraints, problems, and possibilities are recognized, and this will lead to the future development of bitemporal and multitemporal CD approaches and techniques for understanding land cover dynamics [6] .

In 2021 Saurabh Kumar, Shwetank proposed the goal of this research is to create a spectral signature for land use classifications and extract features from it. MTMS (multi-temporal and multi-spectral) Landsat r data images dataset, Saurabh Kumar and Shwetank suggested this work in 2021 to establish a spectral sign and feature extraction of land use classifications. From 2003 to 2017, the imagery dataset collected three photos from the Landsat satellite system's various sensors. The pre-processing of imagery is critical for extracting geographical data and analyzing surface-use features. For different years (2017, 2010, and 2003), the classification accuracy utilizing the ANNs approach was 90.10 percent, 75.75 percent, and 78.37 percent, respectively[7] .

DATA SET

Satellite image Classification Dataset-RSI-CB256[8], This dataset has four different classes mixed from Sensors and google map snapshot. The whole dataset has 5631 images in jpg format and each class has about 1500 images. The training dataset was used to calibrate the chosen model, whereas the testing dataset was used to evaluate the models' performance using a dataset that was not used to train them[9] .

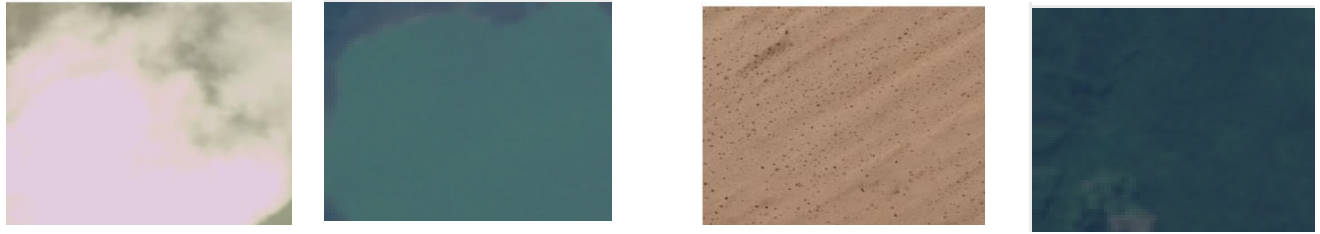


Figure 1. classes of the data set.

TABLE 1. The details of dataset

CLASS	NO. OF PICS
Cloudy	1500 cloud pics at different time
Green area	1500 green area pics at different time
Desert	1131 desert pics at different time
Water	500 water lake pics at different time

PREPROCESSING OF IMAGES

This stage's main purpose is to improve picture data to remove unpleasant distortions or enhance particular elements of images that are needed for further processing[10] . Image enhancement is a technique for removing undesired distortion caused by loss of contrast, unwanted noise, inappropriate intensity saturation, blurring effect[11]. There are several stages for this optimization:

I- COLOR IMAGES TO HSV IMAGES

Color model is a method for defining color based on the three primary features of color: hue, saturation, and brightness. Hue (H) is the most fundamental aspect of color, and it is simply the color. Names like red and yellow come to mind. According to the current state of affairs, the conventional color wheel has a range of 0 to 360 degrees[12]. saturation (vibrancy) and value (brightness) to the transformation from RGB color space to HSV color space is given in the following equations:

$$H = \cos^{-1} \frac{\frac{1}{2}(2R-G-B)}{\sqrt{(R-G)^2 - (R-B)(G-B)}} \quad (1)$$

$$S = \frac{(R,G,B) - (R,B,G)}{\max(R,G,B)} \quad (2)$$

$$V = \max(R,G,B) \quad (3) \quad [13]$$

II- COLOR TO GRAY IMAGES CONVERSION

The majority of the advantages of converting a color image to a grayscale domain include having less data because the grayscale domain has one channel instead of three channels in the RGB domain, which allows for faster processing in other stages (feature extraction and training phase) with minimal influence from the brightness[14]. In many circumstances, it is regarded as a nuisance. Lightness, chroma, and hue are the perceptual qualities that represent color as a three-dimensional phenomenon. As a result, the conversion from color to grayscale is extremely lossy, reducing a 3-D representation to a 1-D representation. The brightness information is kept while the chroma and color information is eliminated in the most usual method. For example, in Postscript, the conversion is as follows:

$$0.30 R + 0.59 G + 0.11 B = GRAY \quad (4) \quad [15]$$

III- HISTOGRAM EQUALIZATION

It's a spatial domain method that produces an output image with a uniform pixel intensity distribution by flattening and extending the histogram of the output image routinely. Because of its simplicity and relative superiority over other traditional methods, this approach is commonly used in the picture enhancement paradigm. The input image histogram is the probability density function (PDF) and cumulative density function (CDF) are calculated using this method (CDF). Replace the gray levels in the input image with the new gray levels using these two functions PDF and CDF, and then generate the processed image and histogram for the resultant image. discovered that the gray level intensities are stretched and lowered systematically when comparing the input image histogram to the processed image histogram[11]. This helps when the satellite images are captured in the dark. The normalized histogram of the intensity image is:

$$pn = \frac{\text{The number of pixels with intensity } n}{\text{The total number of pixels}} \quad (5)$$

Where $n=0,1,2,...L-1$ and L is the total number gray levels in the image i.e., 256. The histogram equalized image is given by

$$g(i,j) = (L-1) \sum_{n=0}^f i_j pn \quad (6) \quad [13]$$

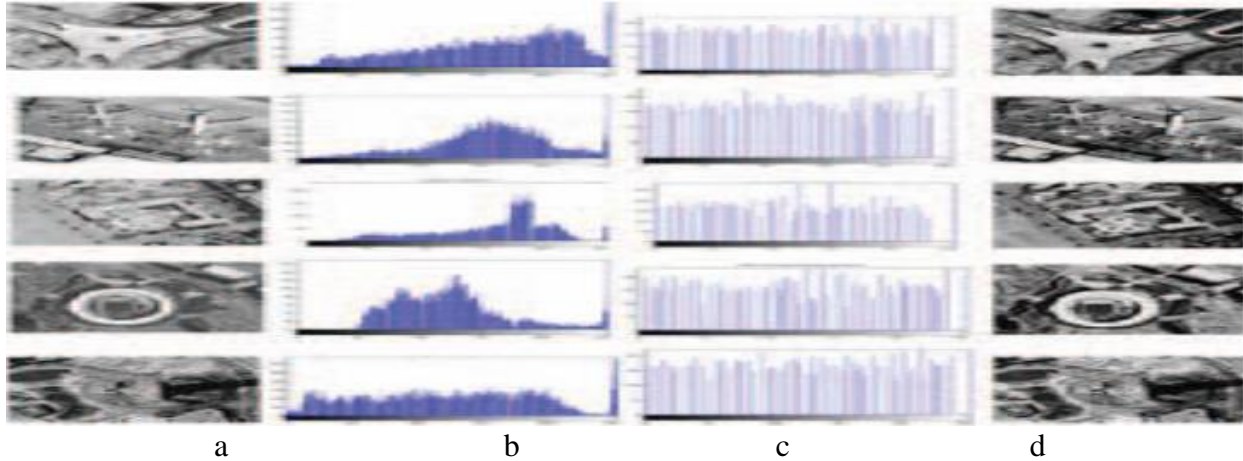


FIGURE 2. Histogram equalization of value component (image) (a) value component (b) histogram of value component (c) histogram equalization (d) histogram equalized image[13].

IV- BLUR THE IMAGE

Its convolution using a low pass filter kernel to achieve blurring. It's useful for noise reduction and smoothing images to remove minor detail texture or noise. It's frequently useful when using image processing algorithms that look at the image's finer details [16]. the degree of blurring. Extract the features of interest using CVIP tools[17] .

VII- RESIZE

Because there are many distinct images, each with a different capture size, the image must be resized to have a unique dimension for feature extraction[14]. the resize algorithm was used to make the features in the image clear and reduce the size to make it more accurate by removing the extra features, so the processing time becomes less.

FEATURE EXTRACTION USING FAST FOURIER TRANSFORM (FFT)

The feature extraction module utilizes the minutiae features extracted from the satellite images recognition as proposed [15]. The reflectance values will be utilized to extract the image's spectral information[18] .In this paper the equation of fast Fourier Transform is used for the extraction of the features. the (FFT) Most transforms have rapid algorithms, and many of them are based on input data with several components that are a power of two, which is frequent for photos. These algorithms, in general, take advantage of the numerous unnecessary calculations required and work to remove them. Computer Vision and Image Processing Tools (CVIPtools) use fast techniques based on powers of two, which means that any image that is not a power of two will be zero-padded [14]. This step's main purpose is to convert M examples from the time domain to the frequency domain. This phase is intended to remove superfluous mathematical computations and allow for the analysis of a signal's spectrum qualities.

$$X_n = \sum_{k=0}^{M-1} (X_k e^{-2k\pi n/M}) \quad (7)$$

$n=0,1,2,3, \dots, M-1$

In a situation of complex only total value is considered. Where frequency range $0 \leq f \leq F/2$ corresponds to $0 \leq n \leq M/2-1$ and frequency range $-F/2 \leq f < 0$ corresponds to $M/2+1 \leq n \leq M-1$. Where F means the sampling frequency[19].

VECTOR QUANTIZATION (VQ)

It's a process for taking a large group of feature vectors and producing a small group of feature vectors that reflect the spreading's centroids, or points spaced so that the average distance between each of the other locations is minimized. Because accumulating each of the feature vectors is inefficiently generated from training utterances, VQ is used.[20] While the VQ technique is time-consuming to compute, it saves time during the testing phase. The Euclidean distance will be given as

$$d(a,b) = \sqrt{\sum_{j=1}^a (a_j - b_j)^2} \quad (8)$$

In where a_j denotes the j th component of the input vector, and b_j denotes the j th element of the code-word b_i [21] .

MACHINE LEARNING ALGORITHMS

Machine learning is a facts analysis technology that automates the building of analytical models. It is also considered a division of Artificial Intelligence (AI) since it is founded on the concept that a machine can learn from data, discover outlines, and make choices with the least amount of human involvement[22]. In this paper, four algorithms which are logistic regression (LR), random forest (RF), stochastic gradient descent (SGD), and Naive Bayes (NB) are used, the following sections (7.1, 7.2, 7.3 and 7.4) explain the details of these algorithms:

I- LOGISTIC REGRESSION (LR)ALGORITHM

A logistic regression model that models the posterior class probabilities $\Pr(G = j | X = x)$ for the J classes is a better method to use regression for classification tasks. Given class probability estimates, logistic regression models these probabilities using linear functions in x while guaranteeing that they total to one and remain in the range $[0, 1]$ [23] . With the following equation, a personality's hazard of an outcome may be calculated using the individual's detected predictor values and the model's intercept and regression constants in a logistic regression-based calculation model.

$$LP = \log \left(\frac{p}{1-p} \right) = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_k X_k \quad (9)$$

$$P = \frac{\exp(lp)}{1 + \exp(lp)} \quad \text{or else} \quad P = \frac{1}{1 + \exp(-lp)}$$

where LP is the linear predictor, p is the prophesied probability and is the interrupt, and I is the predictor i's regression coefficient. Xi is the observed value for predictor i and k is the model's number of predictors[24] .

II- RANDOM FOREST (RF)ALGORITHM

The Random Forest (RF) is a tree-structured hierarchical collection of basic classifiers. Text data often contains a large number of dimensions. There are a lot of useless characteristics in the dataset. For the classifier model, just a few key properties are useful. The RF algorithm selects the most essential relevant feature based on a simple fixed probability. Bierman developed the RF technique by projecting a random sampling of feature subspaces to multiple decision trees using sample data subsets[25] .

III- STOCHASTIC GRADIENT DESCENT (SGD) ALGORITHM

It's a time-saving technique also known as incremental gradient descent It's also a stochastic approximation method that takes the average of previous gradients and proceeds in that direction while decreasing exponentially. As a result, it is an optimization strategy with various advantages, such as providing not just ideal sample complexity but also optimal runtime[26].

VI- NAIVE BAYES (NB) ALGORITHM

The Naive Bayes method is a straightforward probabilistic classifier that creates a set of probabilities by counting the frequency and combinations of values in a batch of data. The Bayes theorem is applied to the method, which assumes that all qualities are equal. Given the value of the class variable, it is independent. This in practice, the premise of conditional independence rarely holds. As a result of this, the term "Nave yet" was coined. In general, the algorithm performs well and learns quickly. A variety of supervised classification issues. The Naive Bayes classifier is a probabilistic classifier that uses the Nave Bayes rule to calculate the latter probability of a character r being in class c [27] .

$$P(c|r) = \underset{P(r)}{\operatorname{argmax}} \frac{P(c)P(r|c)}{P(r)} \quad (10)$$

Where: P (r|c) is the probability of predictor given class, which is the likelihood(C) is the prior probability of a class, which is calculated as

$$P(C)=NCI/N \quad (11)$$

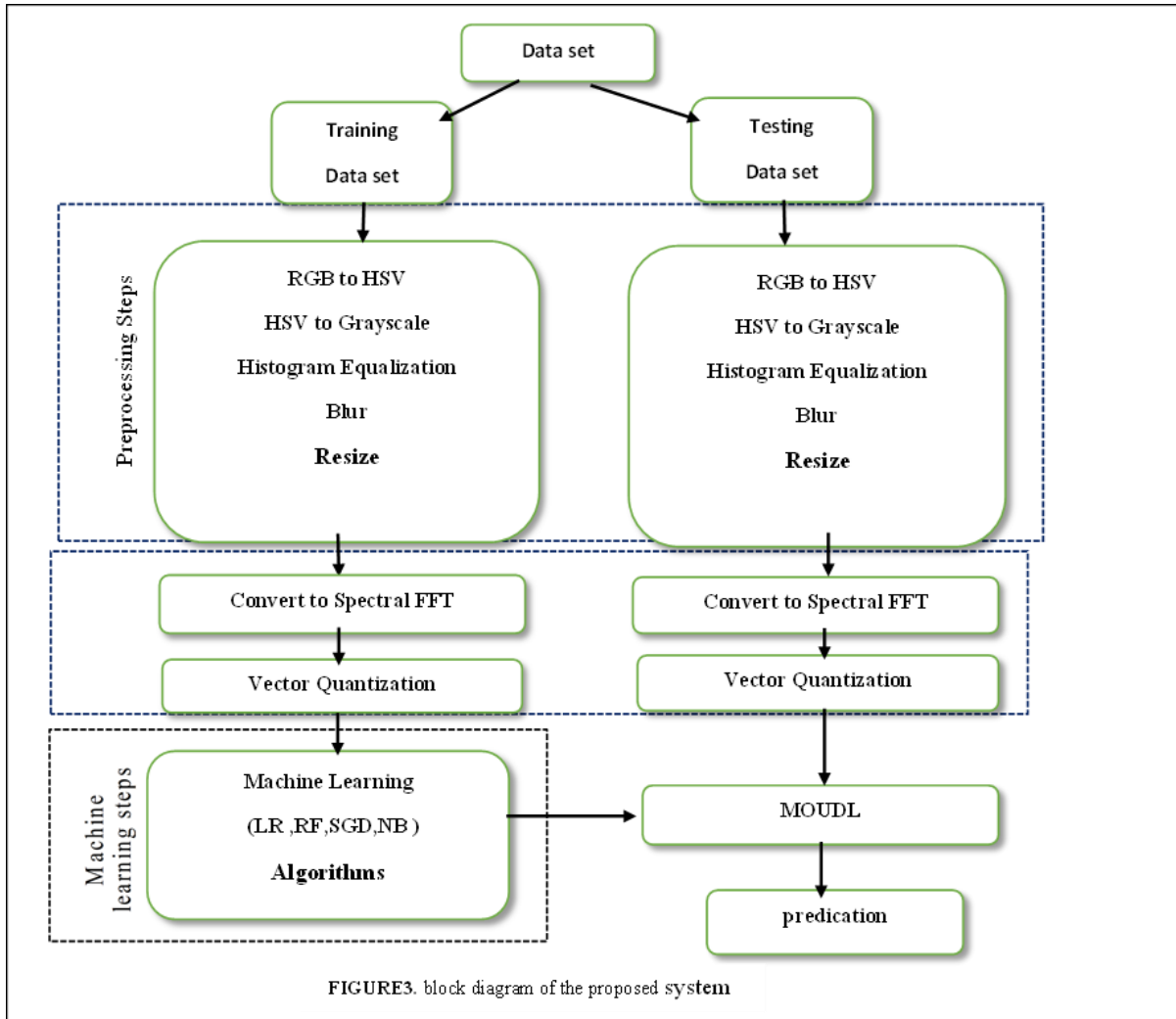
Where: N is the whole number of images in the training set, and NCI is the number of pictures in class CI.

P(r) is the prior probability of analyst (r) for all classes[21].

PROPOSED SYSTEM

Figure (3) refers to the block diagram of the proposed system. In this system Satellite image Classification Dataset contains four different classes of satellite images, The whole dataset has 5631 images in jpg format and each class has about 1500 images. This set is divided into 70% for training and 30%for testing. The proposed system consists of main six steps, Following the capture of the photographs, they were all processed to increase details and prepared for the training portion[28] .Image preprocessing as RGB to HSV and HSV to grayscale and histogram equalization, convert to spectral by (FFT)algorithm, vector quantization, and applied four machine learning algorithms to classify satellite images to its classes.

I- PREPROCESSING



First step: convert satellite images from RGB to HSV as shown in equations (1),(2), and (3) .Then convert these images from HSV to grayscale in equation (4) explain in figure (4)

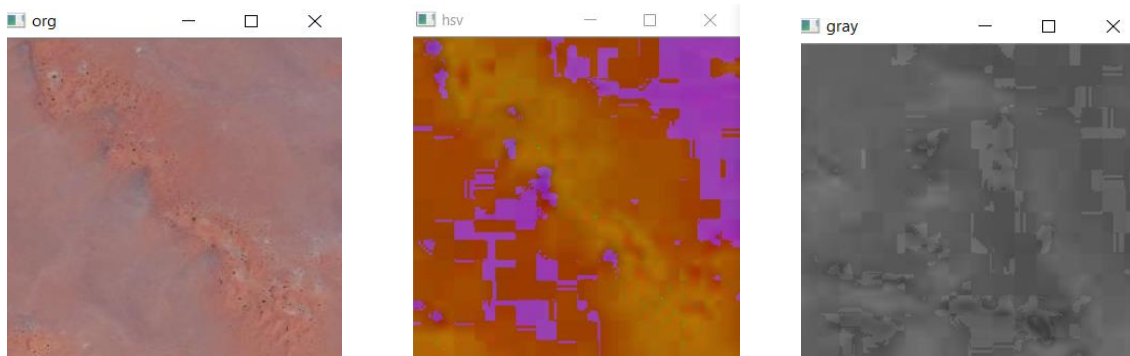


FIGURE 4. Convert from RGB to HSV and HSV to grayscale.

Second step: apply histogram equalization as seen in equation (6) .and media blurs to obtain best images without noise as shown in figure (5) then using resize of images to Lowe dimensions 20×20 for best feature extraction.

FIGURE 5. apply histogram equalization and blur.

II- FEATURE EXTRACTION

First step: this paper used FFT (fast Fourier transform) to feature extraction using the equation (7). That explains the work of converting images data after resizing to its spectral signature as shown in figure (6).

Second step: apply vector quantization in equation (8) It is the process of turning a continuous range of values into a finite range of discrete values, to reduce a group of features and obtain one dimension vector.

III- MACHIN LEARNING

In this step, four algorithms of Machin learning classifiers are applied to the training data set and tested by testing data set. the result of precision, recall, and f1-score by using equations (12),(13), and(14) explain in section 9. the next figure is an algorithm of the proposed system.

Algorithm 0f proposed system
Input: image Classification Dataset-RSI-CB256. Output: Class label
A: No. of images from the data set B: HSV of image C: Grayscale of the image D: Histogram equalization E: A blur of the image F: Resize of image G: Feature extraction FFT H: Vector quantization of features Begin Step 1. the data set divided to 70% of the dataset is for training and 30%for testing in A. Step 2. pre-processing, by applying RGB to HSV in B, HSV to grayscale in C, Histogram equalization in D, blur in E, and resize in F. Step 3. Get spectral by using FFT in equation (7) in G. Step 4. Vector quantization in equation (8) in H. Step 5. The training phase will be done on the training portion of the dataset by trying four types of Classifiers are (SGD, NB, LR, RF) to distinguish between the four classes of satellite images. Step 6. The testing phase will be done on the testing part of the dataset in which the results of the classifiers of the training phase are examined and evaluated. End.

EVALUATING PERFORMANCE MEASURES

Precision, recall, and the f1-score are only a few of the statistical measurements utilized to boost performance power.

Precision is measured by dividing the total number of positive forecasts by the number of actual positive forecasts. A recall is one of the most important metrics in models with unbalanced datasets. The true positive rate is calculated in the model. The F1-score may be viewed as the average of recall and accuracy[29] .

$$Precision = \frac{true\ positive}{true\ positive + false\ positive} \quad (12)$$

$$Recall = \frac{true\ positive}{true\ positive + false\ negative} \quad (13)$$

$$F1 = 2 * \frac{precision * recall}{precision + recall} \quad (14)$$

[30]

I- LOGISTIC REGRESSION (LR)ALGORITHM

According to figure (7). when using the LR classifier with using feature extraction method (FFT) and select 20 features. our proposed system Overall Precision is 74%, recall is 45% and f1-score is 54%.

FIGURE 7. Performance of logistic regression (LR)algorithm.

II- RANDOM FOREST (RF)ALGORITHM

As it can be seen from figure (8); when using the RF classifier and using the feature extraction method (FFT) with 20 features. our proposed system Overall precision is 56 %, recall is 48% and f1-score is 51%.

FIGURE 8. Performance of random forest (RF)algorithm.

III- STOCHASTIC GRADIENT DESCENT (SGD)ALGORITHM

About SGD, the outcomes of this paper were shown in Figure (9). This classifier obtained the highest precision compared with others, in which the precision of Performance is 81%, recall is 42% and f1-score is 55%.

FIGURE9. Performance of Stochastic Gradient Descent (SGD)algorithm.

IV NAIVE BAYES (NB) ALGORITHM

The algorithm performs well and learns quickly, but also the result acquired from it is not optimal. when applying this algorithm as seen in Figure (10),57% the precision, recall is 50% and the f1-score is 52%.

FIGURE10. Performance of Naive Bayes (NB) algorithm.

Based on the testing findings of the four classification methods, it was discovered that the SGD algorithm has the highest precision, as shown in Table (2).

TABLE 2. Comparison between Experiments.

Methods	precision	Recall	F1-score	Accuracy
SGD	81%	45%	56%	45%
LR	74%	45%	54%	45%
NB	57%	50%	52%	50%
RF	56%	48%	51%	48%

As a result, the dataset is split into two parts: training and testing. 70% of the dataset is utilized for training, while the remaining 30% is used to test the classifier. Using the Python language, the program was written, which is a language that allows writing reliable systems that are simple to learn and whose code can be simply understood. Therefore, it is ideal for programming machine learning algorithms.

CONCLUSION

The use of spectral signature to extract features before entering data into machine learning classifiers makes these classifiers able to extract stronger features from the features inside them, and this makes the classification process effective because it obtains high weights for comparison in the testing phase, after applying the FFT equation to transform image data into a signal or Energy then selects the best features and minimizes non-critical features, there was an urgent need to use efficient data mining techniques to improve the big data classification of satellite images. When compared with several classifiers such as Naive Bayes (NB), random forest (RF), Logistic Regression (LR), and Stochastic Gradient Descent (SGD), it provides the best precision. Optimization methods like deep learning may be applied to get the best possible classification and high Precision, recall, accuracy, and the f1-score.

REFERENCES

- [1] Chen Y, Wang Q, Wang Y, Duan S B, Xu M and Li Z L 2016 A Spectral Signature Shape-based Algorithm for Landsat Image Classification *ISPRS Int. J. Geo-Information* **5**
- [2] Yang I and Acharya T D 2015 Exploring Landsat 8 *Int. J. IT, Eng. Appl. Sci. Res.* **4** 2319–4413
- [3] Padma S and Sanjeevi S 2014 Jeffries Matusita based mixed-measure for improved spectral matching in hyperspectral image analysis *Int. J. Appl. Earth Obs. Geoinf.* **32** 138–51
- [4] Bruzzone L and Serpico S B 1997 Detection of changes in remotely-sensed images by the selective use of multi-spectral information *Int. J. Remote Sens.* **18** 3883–8
- [5] Ávila Vélez E F, Escobar Escobar N and Morantes Choconta C F 2019 Applying satellite images to spectral signature development of maize production (*Zea mays* L.) under colombia's middle tropics conditions *Entramado* **15** 256–62
- [6] Panuju D R, Paull D J and Griffin A L 2020 Change detection techniques based on multispectral images for investigating land cover dynamics *Remote Sens.* **12**
- [7] Kumar S, Shwetank S and Jain K 2021 Development of Spectral Signature of Land Cover and Feature Extraction Using Artificial Neural Network Model *Proc. - IEEE 2021 Int. Conf. Comput. Commun. Intell. Syst. ICC CIS 2021* 113–8
- [8] Anon satellite-image-classification @ www.kaggle.com
- [9] Mohammed Z A, Abdullah M N and Al-hussaini I H 2021 Predicting Incident Duration Based on Machine Learning Methods *Iraqi J. Comput. Commun. Control Syst. Eng.* 1–15
- [10] Maknun C L, Rosjanuardi R and Jupri A 2018 Lesson Design on the Relationship Between Radian and Degree *AIP Conf. Proc.* **2014**
- [11] Singh R P and Dixit M 2015 Histogram Equalization: A Strong Technique for Image Enhancement *Int. J. Signal Process. Image Process. Pattern Recognit.* **8** 345–52
- [12] Li S and Guo G 2010 The Application of Improved HSV Color Space Model in Image Processing *Proc. 2010 2nd Int. Conf. Futur. Comput. Commun. ICFCC 2010* **2** 10–3
- [13] Ganesan P D V R 2013 VALUE BASED SEMI AUTOMATIC SEGMENTAION OF SATELLITE IMAGES USING HSV COLOR SPACE, HISTOGRAM EQUALIZATIONAND MODIFIED FCM CLUSTERING ALGORITHM 2013 *International Conference on Green Computing, Communication and Conservation of Energy (ICGCE)* vol 8 pp 77–82
- [14] Aliris I S A-J 2015 Face Recognition in Digital Video by Using Artificial Intelligence 7
- [15] Bala R and Braun K M 2003 Color-to-grayscale conversion to maintain discriminability *Color Imaging IX Process. Hardcopy, Appl.* **5293** 196
- [16] Zhang W, Quan W and Guo L 2012 Blurred Star Image Processing for Star Sensors Under Dynamic Conditions *Sensors (Switzerland)* **12** 6712–26
- [17] Horace H-S I 1990 *Digital Image Processing and Computer Vision* vol 8
- [18] Jasim O, Hasoon K and Sadiq N 2019 Mapping LCLU Using Python Scripting *Eng. Technol. J.* **37** 140–7
- [19] Bharti R and Bansal P 2015 Real Time Speaker Recognition System using MFCC and Vector Quantization Technique *Int. J. Comput. Appl.* **117** 25–31
- [20] Soong F K, Rosenberg A E, Rabiner L R and Juang B H 1985 Vector Quantization Approach To Speaker Recognition. *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.* 387–90
- [21] Mohammed R A and Hassan N F 2020 *Proposal Personal Identification Model based on Multi Biometric Features*
- [22] Li L, Zhang Y and Zhao Y 2008 K-Nearest Neighbors for Automated Classification of Celestial Objects *Sci. China, Ser. G Physics, Mech. Astron.* **51** 916–22
- [23] Landwehr N, Hall M and Frank E 2005 Logistic Model Trees *Mach. Learn.* **59** 161–205

- [24] Takada T, Hoogland J, van Lieshout C, Schuit E, Collins G S, Moons K G M and Reitsma J B 2022 Accuracy of Approximations to Recover Incompletely Reported Logistic Regression Models Depended on Other Available Information *J. Clin. Epidemiol.* **143** 81–90
- [25] Sumathi S and Pugalandhi G K 2021 Cognition based spam mail text analysis using combined approach of deep neural network classifier and random forest *J. Ambient Intell. Humaniz. Comput.* **12** 5721–31
- [26] Alaoui S S, Farhaoui Y and Aksasse B 2017 Classification Algorithms in Data Mining – A Survey *A Comp. Study Classif. Tech. Data Min. Algorithms* **6** 1–6
- [27] Yildirim P 2015 Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease *Int. J. Mach. Learn. Comput.* **5** 258–63
- [28] Abed I 2019 Lung Cancer Detection from X-ray images by combined Backpropagation Neural Network and PCA *Eng. Technol. J.* **37** 166–71
- [29] Marza N H 2021 Classification of Spam Emails using Deep learning 63–8
- [30] Mahmoud M M and Nasser A R 2021 Dual Architecture Deep Learning Based Object Detection System for Autonomous Driving *Iraqi J. Comput. Commun. Control Syst. Eng* **21** 36–43