# IndiaSat: A Pixel-Level Dataset for Land-Cover Classification on Three Satellite Systems - Landsat-7, Landsat-8, and Sentinel-2


Chahat Bansal
Indian Institute of Technology
Delhi, India
chahat.bansal@cse.iitd.ac.in

Hari Om Ahlawat
Indian Institute of Technology
Delhi, India
mcs182093@iitd.ac.in

Mayank Jain
Indian Institute of Technology
Delhi, India
mcs182011@iitd.ac.in

Om Prakash
Indian Institute of Technology
Delhi, India
om.prakash.mcs19@cse.iitd.ac.in

Shivani A Mehta
Indian Institute of Technology
Delhi, India
Shivani.A.Mehta@cse.iitd.ac.in

Deepanshu Singh
Indian Institute of Technology
Delhi, India
Deepanshu.Singh.cs118@cse.iitd.ac.in

Harshavardhansushil Baheti
Indian Institute of Technology
Delhi, India
cs5180407@cse.iitd.ac.in

Suyash Singh
Indian Institute of Technology
Delhi, India
cs5180646@cse.iitd.ac.in

Aaditeshwar Seth
Indian Institute of Technology
Delhi, India
aseth@cse.iitd.ernet.in



## ABSTRACT

Land-cover (LC) classification is required for land management and planning models, and is increasingly done through remote sensing data. Supervised machine learning methods applied to satellite imagery can help with high-resolution LC classification but demand a labeled dataset for training and evaluation of the models. The availability of such datasets is limited though, especially for developing regions like in India. We describe a large pixel-level dataset, *IndiaSat*, that we have curated and provided for open use, consisting of 180,414 pixels labeled into four LC classes: greenery, water bodies, barren land, and built-up area. Initial labels are obtained through the crowd-sourced mapping platform Open Street Maps (OSM), and then manually curated and corrected. We describe our data cleaning methodology and ensure spatial diversity across different geographic regions in the country. We show that the *IndiaSat* dataset can be used to train simple classifiers deployed on commodity platforms like Google Earth Engine (GEE) for three popular and openly accessible satellite systems: Landsat-7, Landsat-8, and Sentinel-2, with high accuracy, and to additionally build LC change detection models to determine pixel-level changes over a sequence of several years.



## CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Applied computing** → *Computing in government*.




*COMPASS '21, June 28-July 2, 2021, Virtual Event, Australia*



ACM ISBN 978-1-4503-8453-7/21/06…$15.00
https://doi.org/10.1145/3460112.3471953

## KEYWORDS

LULC Mapping, Satellite Imagery, Open Street Maps

**ACM Reference Format:**
Chahat Bansal, Hari Om Ahlawat, Mayank Jain, Om Prakash, Shivani A Mehta, Deepanshu Singh, Harshavardhansushil Baheti, Suyash Singh, and Aaditeshwar Seth. 2021. IndiaSat: A Pixel-Level Dataset for Land-Cover Classification on Three Satellite Systems - Landsat-7, Landsat-8, and Sentinel-2. In *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS) (COMPASS '21), June 28-July 2, 2021, Virtual Event, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3460112.3471953


## 1 INTRODUCTION

Land-cover (LC) classification that indicates different features on the Earth's surface, like forests, rivers, croplands, or buildings, is used actively for monitoring and planning of land-use. For example, tracking of anthropogenic activities such as environmental degradation through desertification, loss of farmlands, and the impact of floods [9, 25, 34], are done through satellite maps annotated with LC categories [29], and used by development practitioners to design sustainable land management policies [6].

With the advent of remote sensing and image processing techniques driven by machine learning, LC classification has become more feasible and accurate at fine geographic scales [7]. LC classification broadly falls into two methods: pixel-based and object-based classification [8, 26]. Pixel-based methods use spectral signatures to independently classify each pixel of satellite imagery, while object-based methods use object detection followed by classification to mark contiguous regions as belonging to different LC categories. In this paper, our focus is on per-pixel supervised learning techniques for LC classification, given the lower computational requirements for pixel-based methods.

LC classification requires reliable ground-truth datasets to build learning systems [16, 19]. Prior work has shown that supervised classifiers for LC mapping often do not generalize across different regions [17], and therefore new datasets are needed to train LC classifiers for new regions. We present a large dataset named *IndiaSat*, of 180,414 pixels, for LC classification in India into four classes:

Greenery, water bodies, barren land, and built-up areas. The dataset incorporates spatial diversity across different agro-climatic zones in India and classifiers trained on this data perform robustly in rural and urban settings as well. The dataset is curated for the year 2019 and can support the training of three satellite systems active during this year: Landsat-7, Landsat-8, and Sentinel-2. In this paper, we outline a step-by-step process of how we curated this dataset, by improving upon annotated data available from the Open Street Maps (OSM) platform, and we make the dataset available to build LC classifiers on the Google Earth Engine (GEE) platform.

We demonstrate the effectiveness of the dataset by building machine learning models for cross-sectional LC classification as well as to detect longitudinal changes in land classification over multiple years. We achieve a 5-fold cross-validation accuracy of over 98% for all three satellites, for LC classification into four categories. To the best of our knowledge, this is the first publicly available dataset for India for LC classification into four categories. A smaller dataset for classification into two categories of BU (built-up) and NBU (non built-up) was curated by Goldblatt et al. [15], and we show that models built on our dataset are able to outperform models built on this dataset by 7.6% for a two-class classification. We further build and evaluate a method for detecting change over time in land-cover on a per-pixel basis [3, 15], and are able to achieve a performance of over 85% in LC change detection over a period of four years. Additionally, we use heuristics to make the classification robust against seasonal changes in land cover, such as agricultural land appearing as barren during the summer months [1]. Our entire codebase for LC classification along with the *IndiaSat* dataset is available for public use on Github: https://github.com/ICTD-IITD/IndiaSAT.

## 2 RELATED WORK

The extensive literature for LC classification can be grouped broadly into pixel-based and object-based methods. Both these approaches have proven to perform well in identifying generalized land cover classes such as built-up, water, and green land [20, 37]. Pixel-based methods that do not use spatial information of neighbouring pixels can have a lower accuracy than object-based methods though [10, 24], but they are also computationally lower cost to execute and platforms such as Google Earth Engine (GEE) provide free access to several satellite systems on which pixel-based methods can be easily applied [11, 27]. In addition, as pointed out by Huth et al., object-based methods may not be suitable in certain regions that have small-sized objects such as residential neighbourhoods having small parks, which is specifically the case in many developing country regions, and pixel-based methods may be better suited here [20].

Supervised machine learning tasks for pixel-level classification require high-quality labeled datasets that are sufficiently large to capture spatial variations and have a balanced class representation [23]. Creation of labeled datasets for satellite imagery has seen different methods in India. Roy et al. provided land-use mapping into nineteen classes for India, for the years 1985, 1995, and 2005, using the Landsat-4 and Landsat-5 satellite data systems [28]. The data was manually curated for 2005, but no validation was conducted for other years. Later in 2016, Goldblatt et al. released a pixel-level dataset for built-up/non-built-up mapping on Landsat-7 and Landsat-8 [16]. This dataset comprised of manually labeled 21,030 polygons (30m resolution), sampled from different parts of the country. The manual process of data annotation followed by the authors was an arduous and time-consuming activity. We used annotated data from Open Street Maps (OSM) as our starting point, and checked and corrected the data manually to generate a dataset almost 10x in size as the Goldblatt dataset, more detailed with four LC classes instead of just two, spatially more diverse, and are able to deliver better performance on models trained using this data. Since OSM data can be noisy [35], we also describe several steps we undertook to improve the data, and these insights can be useful for other researchers who may want to build LC datasets for different regions.

## 3 DATASET

We outline in detail the step-by-step methodology we followed to create training and out-of-sample testing datasets for pixel-based LC classification.

### 3.1 Preparation of the IndiaSat dataset

We choose the following four categories for LC classification, as being some of the prominent ones useful for several applications:

(1) *Greenery:* This represents vegetative lands like parks, forests, and agricultural lands, often monitored for environmental policies related to green-cover [36].
(2) *Water body:* This includes rivers, lakes, and seas, which are tracked especially for water conservation [13].
(3) *Barren land:* These are non-vegetative lands such as deserts and rocky terrains, helpful to identify regions facing land degradation or identify non-agricultural land for urban expansion [12].
(4) *Built-up (BU):* This denotes man-made constructions such as buildings and roads to identify human settlements, and are actively used to track urbanization dynamics [4].

*Using OSM data to extract relevant polygons against the target LC classes.* The large scale of community-driven geodata available from OSM has lately seen a lot of attention. Even though there are reservations about its accuracy and completeness [30], it has been actively used as a guiding dataset to create labeled data [21]. We started with identifying places in India tagged in OSM against multiple categories, and mapped them to our target LC classes as shown in Table 1. All classes had straightforward equivalent categories in OSM and we were able to obtain corresponding polygon regions from which we could sample pixels for the corresponding LC classes, as shown in Figure 1. However, the polygon markings were some times not accurate, and we next describe different steps we had to undertake to clean the OSM data.

*Noise removal in OSM classes.* We obtained OSM data for the four LC classes in the form of polygon shape-files. A closer examination however against satellite imagery for the year 2019 revealed the following inconsistencies (in the order of their processing)-

(1) *Mixed-polygon annotations:* Sometimes polygons for different classes overlapped with one another. For example, residential areas along a river caused an overlap between the built-up and water-body classes (Figure 2a), or residential areas in rural parts of Orissa overlapped with greenery (Figure 2b). We

| LC class | OSM place type |
|---|---|
| Greenery | forest, nature reserve, park, farms, meadow, orchard, and scrubs |
| Water body | rivers, water reservoirs and natural water bodies |
| Barren land | non-vegetative forest, park, farms, scrubs, cliffs, quarry, and rocky terrain |
| Built-up | buildings |

**Table (1) LC class mapping to OSM place types**

**(a) Greenery** **(b) Water body** **(c) Barren land** **(d) Built-up**

**Figure (1) Sample images of OSM polygon markings for all 4 LC classes included in the IndiaSat dataset**

automatically discarded such OSM polygons that overlapped with polygons of other classes.

(2) *Overlapping polygons:* Multiple annotations for the same place can lead to overlapping polygons, as can be seen in Figure 2c. Such selections can lead to sampling of duplicate pixels. Only one polygon was randomly selected out of such overlapping polygons.

(3) *Incorrect boundary delineations:* Some OSM polygons did not have accurate boundaries. For example, polygons marked as parks, as shown in Figure 2d, may extend into residential buildings. Similarly, the boundaries of water bodies may change in different seasons, as in Figure 2e, but a larger area may be marked in OSM. We had to manually inspect and discard such polygons which had ambiguous boundaries.

(4) *Delineation between greenery and barren lands:* Manual inspection in the previous step revealed that simply mapping certain OSM categories representative of greenery may not always be accurate. For example, OSM places marked as forest areas in the state of Orissa are lush green with vegetation (Figure 2f), while the forest areas marked in Chattisgarh (Figure 2g) and Madhya Pradesh (Figure 2h) include large patches of rocky terrains and desert-like features that should be categorized as barren-lands. We therefore carefully visualized each polygon against the satellite imagery for the year 2019 to check whether it should be assigned to a particular LC class or dropped, and whether pixels sampled randomly from the polygons can be unambiguously mapped to the corresponding LC class. We used Google Earth Pro to do this check, and in some cases we used a modified polygon marking to ensure consistency.

The last two manual steps for data curation were time-consuming but an essential activity. The entire data cleaning process for IndiaSat took 3-4 months by a group of three students devoting a few hours on most days towards this activity. However, given the experience gained through this study of the kind of noise to expect, we feel that following the above mentioned steps to create a similar dataset for a new region should be possible in two to three weeks.

*Ensuring spatial diversity.* India has significant diversity across its different geographic regions, and we ensured that we sampled polygons from different agro-climatic zones in the country. The country is divided into 15 agro-climatic zones based on features like soil type, temperature, rainfall and water resources. We sampled the dataset from 12 zones out of these 15, leaving aside zones 1 and 2 which include the Himalayan ranges and zone 15 which includes island regions. Owing to the storage and computation limitations of the Google Earth Engine platform, we also had to be sensitive to not create too large a dataset. Eventually, we selected 1,855 polygons from which we extracted 180,414 pixels, as shown in Table 2. These pixels were distributed across the different agro-climatic zones, as detailed in Table 3 and shown in Figure 3. The dataset currently doesn't contain enough polygons of type greenery from zones 3, 12, and 13, which can be taken up in future work. We call this labeled data *IndiaSat* which we used to train LC classifiers, as described in the next section.

| LC class | #Polygons | #Pixels |
|---|---|---|
| Greenery | 92 | 72,492 |
| Water bodies | 112 | 60,275 |
| Barren-land | 274 | 42,473 |
| Built-up areas | 1377 | 5174 |

**Table (2) Size of IndiaSat for different LC classes.**

## 3.2 Preparation of an out-of-sample test dataset

To further check the accuracy of classifiers we train on the *IndiaSat* dataset, we additionally constructed two out-of-sample datasets.

In the first dataset, we manually marked four LC classes in five cities from different parts of the country. This was also done with

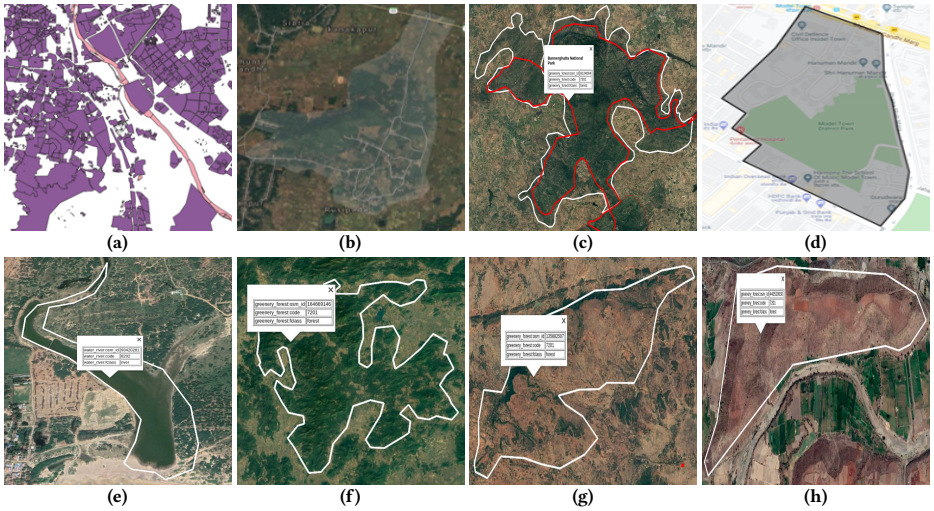



**Figure (2) Snapshots of OSM polygon annotations for (a) Residential areas on sides of a river, (b) Residential area in Orrisa, (c) Overlap of forest shapes, (d) Wrongly marked boundary for a park, (e) Shrunken boundary of water body, (f) Forest in Orissa, (g) Forest in Chattisgarh, and (h) Forest in Madhya Pradesh.**

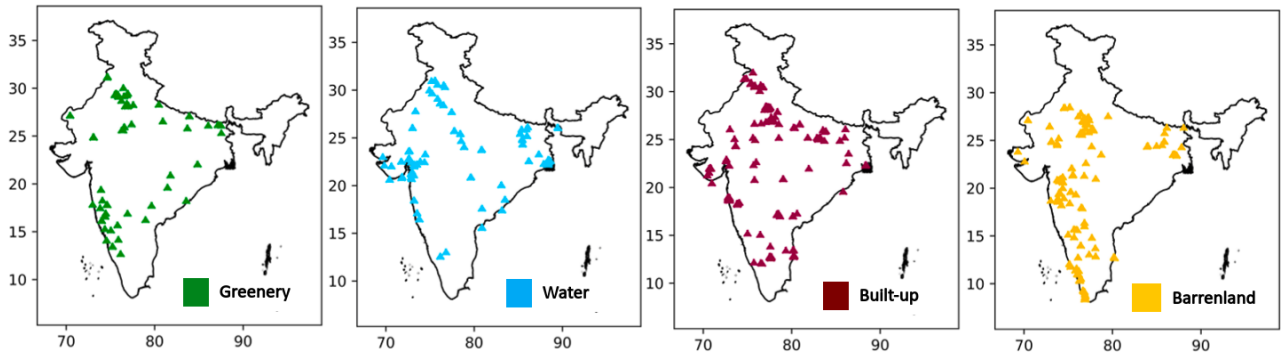



**Figure (3) Distribution of LC polygons from different agro-climatic zones in India.**

| | Greenery | Water body | Barren-land | Built-up |
|---|---|---|---|---|
| Zone-3 | - | 9936 | 14999 | 83 |
| Zone-4 | 164 | 1016 | 2629 | 50 |
| Zone-5 | 5391 | 64 | 18 | 650 |
| Zone-6 | 35338 | 1180 | 243 | 1200 |
| Zone-7 | 1893 | 6291 | 1081 | 14 |
| Zone-8 | 252 | 7507 | 1148 | 44 |
| Zone-9 | 8199 | 10463 | 7290 | 1300 |
| Zone-10 | 7016 | 1612 | 3047 | 1045 |
| Zone-11 | 1598 | 808 | 183 | 21 |
| Zone-12 | - | 8586 | 845 | 593 |
| Zone-13 | - | 12412 | 1946 | 126 |
| Zone-14 | 12641 | 400 | 9044 | 48 |

**Table (3) Distribution of pixels from different agro-climatic zones in India.**

a view to test the robustness of the learned models specifically for urban settings. A team of five students working full-time for approximately one week marked the different LC classes using the Google Earth Pro software for the year 2018. The details of this dataset are given in Table 4a.

With the second test dataset, we additionally wanted to evaluate whether models built on the *IndiaSat* dataset can help detect change that has happened over a period of several years. Detecting change between 4 LC classes would have required 16 combinations. We therefore manually built this dataset with three types of pixel annotations: Constantly built-up (CBU), Constantly non-built-up (CNBU), and Changed. This was constructed to classify change during the period 2016-2019 for which all the three satellite systems were operational. Here, CBU denotes pixels with man-made construction during all these years, CNBU are the pixels that remain non-built-up throughout, and Changing pixels are the ones which converted from NBU to BU during this period. We assume BU to NBU transitions as rare and possibly anomalous, and we were not able to find any such instances either while building the dataset. We took care to sample CNBU pixels from among different LC classes to capture transitions from greenery to built-up, barrenland to built-up, and also water-bodies to built-up. The details of this dataset are shown in Table 4b.

| City | #Greenery Pixels | #Water Pixels | #Barren Pixels | #Built-up Pixels |
|---|---|---|---|---|
| Bangalore | 490 | 701 | 416 | 402 |
| Delhi | 342 | 214 | 297 | 266 |
| Gurgaon | 800 | 464 | 598 | 598 |
| Hyderabad | 305 | 291 | 200 | 424 |
| Mumbai | 366 | 538 | 339 | 656 |

(a)

| City | #CBU Pixels | #CNBU Pixels | #Changing Pixels |
|---|---|---|---|
| Bangalore | 195 | 215 | 141 |
| Delhi | 337 | 736 | 364 |
| Gurgaon | 419 | 503 | 414 |
| Hyderabad | 184 | 317 | 101 |
| Mumbai | 304 | 498 | 138 |

(b)

**Table (4) Out-of-sample test dataset description- (a) Test dataset to evaluate 4-class classification for the year 2018, and (b) Test dataset for temporal correction and change classification from year 2016 to 2019.**

# 4 METHODOLOGY

## 4.1 Feature selection

The *IndiaSat* dataset is constructed for the year 2019, and can be used to build models for three satellite systems that were active in this year: Landsat-7, Landsat-8, and Sentinel-2. All three satellite systems provide a uniform coverage of the Earth's surface and report spectral bands at a spatial resolution varying from 10m to 100m. The bands include visible (RGB), near-infrared (NIR), shortwave infrared (SWIR), and some others useful related to aerosols and cloud cover [14, 32, 33]. The GIS community has further developed several derived bands using a combination of these basic bands, such as Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), Normalized Difference Builtup Index (NDBI), and Normalized Difference Moisture Index (NDMI) [22, 31]. Feature vectors for the models are constructed using both the basic bands and derived bands. The minimum, maximum, and median values over the year, of all these bands, are used as features. The four classes in the *IndiaSat* dataset are balanced using Synthetic Minority Oversampling Technique (SMOTE) [5]. A Random Forests (RF) classifier is used on Google Earth Engine to build the classification models. The number of trees is set to 100 for which the best 5-fold cross validation accuracies were obtained. A feature selection process is conducted for each satellite system, starting with using all the bands for the classifier and then removing them one by one. The final features used for the three satellite systems are as follows-

- Landsat-7: RGB, NIR, SWIR, Thermal, NDVI, NDWI, and NDBI
- Landsat-8: RGB, NIR, SWIR, Coastal aerosol, NDVI, NDWI, NDBI, and NDMI
- Sentinel-2: RGB, NIR, SWIR, Red edge, NDVI, NDWI, and NDBI

We next discuss two correction methods we developed, to improve the accuracy of predictions made by the trained models.

**Figure (4) Variation in the spectral signature of the same agricultural land at different times of the year due to the seasonal nature of farming activity.**

## 4.2 Seasonal correction

Several classes such as greenery and water bodies can undergo seasonal changes through the year. For example, Figure 4 shows three satellite images, the first having been created using the first six months of the year (January-June), the second using the last six months of the year (July-December), and the third using all the twelve months. These images produce different classifications. To correct for this, we developed a rule-based procedure as explained below and shown in Figure 5. We take the same three images and feed them to the RF classifier. We then use the following set of rules-

- *Rule 1:* If the same LC class is predicted for a pixel in all 3 images, then that is predicted as the final label for the pixel.
- *Rule 2:* If across all three images, there exist only two distinct outputs, then the final label is selected on a majority basis.
- *Rule 3:* For three distinct outputs across all three images, we select the final label based on its likelihood of being predicted correctly. The non-overlapping spectral signatures of greenery and water-bodies make them more distinctive classes than barren lands and built-up areas [2]. This can also be noticed in a TSNE plot of the feature vectors of the pixels, shown in Figure 6. Therefore, we assign the final label in the following priority order - greenery if any of the three

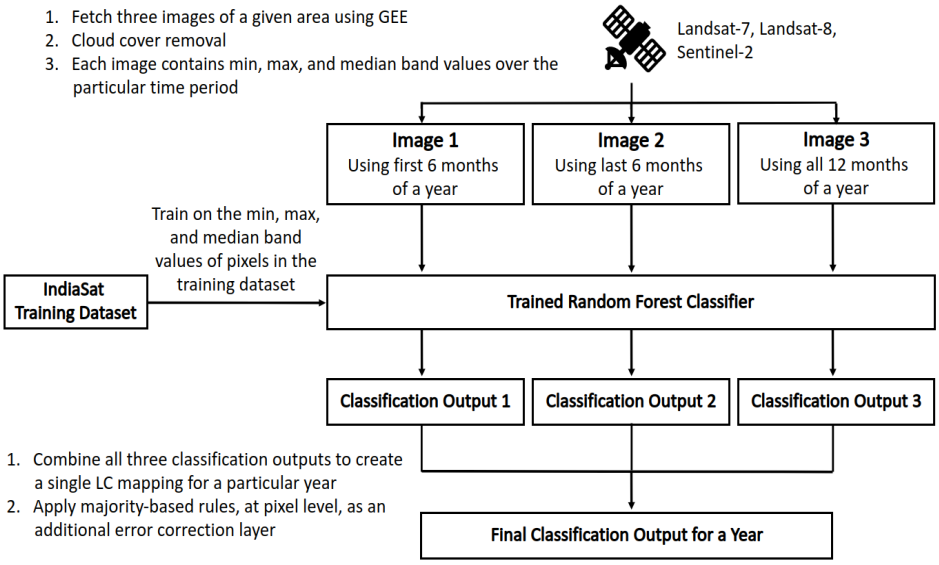



**Figure (5) Methodology for seasonal correction in LC Classification**

labels are predicted as greenery, else water-body if any of the three labels indicate water, else the majority is picked between labels for barren land or built-up area.

These rules essentially act as an ensemble technique that combines several predictions to produce a final prediction.

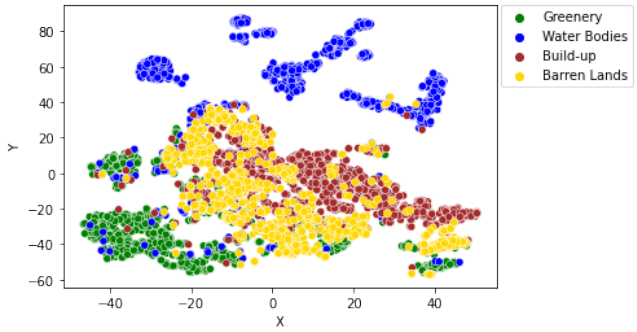



**Figure (6) 2-Dimensional TSNE plot over the spectral signatures of all 4 LC classes for the Sentinel-2 satellite.**

## 4.3 Temporal correction and change classification

Inaccuracies may still emerge in some years due to sensor faults or calibration errors. When operating with the reduced set of two classes, built-up and non-built-up, these inaccuracies may appear as a pixel getting classified as built-up in most years and non-built-up in some year, or vice versa. While some of these cases may reflect legitimate changes that have actually happened, some patterns may appear strange such as built-up areas getting classified as non-built-up in subsequent years, given that demolition activities are rare occurrences. Over a particular range of years, we identify any BU $\rightarrow$ NBU patterns and try to correct them based on the type of NBU subclasses (greenery, water-body, or barren land) that are predicted by the 4-class classifier as shown below:

- *Unique NBU class in the range:* After a pixel is first predicted as BU (from left to right), the percentage of NBU predictions is determined for the successive years. If this percentage is higher than a set threshold, the BU predictions are identified as anomalous, and corrected to NBU. Else, all the NBU predictions for the successive years are corrected to BU. The following examples elaborate this further-
  (1) Example 1- A prediction of $\text{NBU}_{barren} \rightarrow \text{BU} \rightarrow \text{NBU}_{barren} \rightarrow \text{BU} \rightarrow \text{BU}$ over five years, is identified as anomalous in the third year and corrected based on the percentage of NBU predictions after the second year. Since this percentage is less than 50% (threshold for barren land), the prediction for the third year will be corrected to BU.
  (2) Example 2- A prediction of $\text{NBU}_{barren} \rightarrow \text{BU} \rightarrow \text{NBU}_{barren} \rightarrow \text{NBU}_{barren} \rightarrow \text{NBU}_{barren}$ on the other hand, does not lead to any correction for the third year, but the second year is corrected to $\text{NBU}_{barren}$ this time.
  The threshold percentages are decided experimentally for each subclass of NBU (25% for both greenery and water, while 50% for barren land).
- *More than 1 unique NBU class in the range:* In the entire range of years under consideration, first the dominating NBU subclass is identified on a majority basis. In the case of no clear majority, the priority is first given to greenery, then water, and finally to barren lands. All NBU predictions are then considered to be of this dominating subclass and the correction method as outlined above for a *unique NBU class* is applied. For instance, in a prediction of $\text{NBU}_{greenery} \rightarrow \text{BU} \rightarrow \text{NBU}_{barren} \rightarrow \text{NBU}_{water} \rightarrow \text{NBU}_{greenery}$ over five years, greenery becomes the dominating subclass and the second year is corrected to NBU.

This method may not work if a BU → NBU transition is noticed at the end of a range of years though, as with a prediction of BU → BU → BU → BU → NBU. In this case, a majority of the previous years is used to determine the correction, and the fifth year is corrected to BU. If there is no clear majority then if possible we examine a year prior to the range for which change detection is needed.

The corrected sequence of predictions is then used to detect changes in built-up and non-built-up LC classification over a sequence of years. We simply look at the change between the last year and the first year in a given corrected sequence, to produce a three class classification into Constantly Built-up (CBU), Constantly Non-built-up (CNBU), and Changed. While Landsat-7 has been operational since 1999, Landsat-8 since 2013, and Sentinel-2 since 2015, we build a ground truth to evaluate the change classification over the period 2016-2019, keeping the year 2015 as a buffer year for correcting BU pixels when we may need to refer to a prior year. In the future, we aim to create a longer duration ground truth for Landsat-7 and Landsat-8 systems, and experiment with other correction methods similar to the above but applied in a moving window fashion for a longer range of years.

# 5 RESULTS AND ANALYSIS

We next present an evaluation of the models trained on the *IndiaSat* dataset.

## 5.1 Cross-sectional classification

We first report in Table 5 the 5-fold cross validation accuracy to assess the performance of the RF classifier, for all three satellite systems. A high classification accuracy of over 98% is achieved for all the satellites, for a 4-class LC classification. Next, we assess the classifier's performance against the out-of-sample ground truth described in Table 4a. Table 6 presents the results for the Sentinel-2 satellite system across all 5 cities and Figure 7 shows the corresponding prediction maps. Greenery and water-bodies are typically predicted slightly better than the other classes. Barren lands, however, due to their overlapping spectral signature with greenery and built-up class were observed to get misclassified more often. We found that lands with recent dried out vegetation tend to get misclassified as greenery as shown in Figure 8a. Similarly, quarry regions and other rocky terrains, as shown in Figure 8b, get classified as built-up by the classifier due to their similarity with construction material. For this reason, across all cities, we observe a low recall value for barren lands and low precision values for greenery and built-up. We therefore consider an easier problem and use our 4-class classifier to predict two LC classes- built-up (BU) and non-built-up (NBU) and evaluate the classifier's performance.

*Comparative performance.* To test the performance of our classifiers to predict the classes of Built-up (BU) and Non-built-up (NBU), we reduced our 4-class test dataset described in Table 6 to these two classes. Greenery, water-bodies, and barren-land constitute the NBU class, while the built-up class remained the same. For a baseline comparison, we used another dataset of 21,030 polygons by Goldblatt et al., for the year 2014, with the same two LC classes of built-up and non-built-up [16]. This served to conduct a direct comparison of our datasets and models with those by Goldblatt et al. They use an RF classifier with 10 trees, trained on this dataset using all the bands of Landsat-8. Table 7 presents the results for three models-

- *Model-1:* RF classifier with 10 trees trained on Goldblatt's dataset. This acts as a baseline for comparison.
- *Model-2:* RF classfier with 10 trees trained on the *IndiaSat* dataset. This helps evaluate the performance impact of the *IndiaSat* dataset.
- *Model-3:* RF classifier with 100 trees, augmented with seasonal correction, and trained on the *IndiaSat* dataset. Compared with Model-2, this helps evaluate the impact of the model improvements, and compared with Model-1, this helps evaluate the overall improvement we were able to achieve.

On average across all cities, Model-2 gives an improvement of 4.25% over model-1 highlighting the advantage of a larger and possibly more diverse training dataset. Model-3 improves the classification accuracy further by 3.2% over Model-2, and outperforms the baseline model and dataset by 7.6%. The improvements are consistent across all the cities for which the out-of-sample groundtruth was constructed.

## 5.2 Change classification

We next evaluate change classification accuracy between 2016-2019, for all the three satellite systems, based on the methodology described in the previous section. Table 7 gives the macro average precision and recall scores. For most cities, Sentinel-2 which is the newest satellite and provides data at the highest resolution, is able to give a performance close to 90%. Landsat-7 which goes furthest back in time is also able to give a performance of 80%. These results are encouraging and demonstrate the utility of the *IndiaSat* dataset to train simple RF classifiers and yet achieve reasonable performance for LC classification and changes over time.

# 6 CONCLUSION AND DISCUSSION

In this paper, we described our process to create a large labeled dataset for LC classification in India through satellite imagery. We described our data cleaning and curation methodology, and showed that even simple machine learning models deployed on commodity platforms like Google Earth Engine, when trained on this data are able to perform well and better than other currently available models. We also outlined several error-correcting methods we developed to address misclassifications due to seasonal variation in land-use or temporal anomalies due to satellite data reporting errors. We believe that the *IndiaSat* dataset can now be used to train better classifiers to improve the performance further. These models can also benefit other researchers to build LC classification applications. IndiaSat was used to understand urbanization patterns in India [4]. Another application of IndiaSat can also be to verify crowd-sourced annotation of other datasets, as proposed by Helber et al. using the EuroSAT dataset [18].

# REFERENCES

[1] Shivani Agarwal and Harini Nagendra. 2020. Classification of Indian cities using Google Earth Engine. *Journal of Land Use Science* (2020), 1–15.

[2] Abd As-syakur, I Adnyana, I Wayan Arthana, I Wayan Nuarsa, et al. 2012. Enhanced built-up and bareness index (EBBI) for mapping built-up and bare land in an urban area. *Remote Sensing* 4, 10 (2012), 2957–2970.

| LC class | Precision | Recall |
|---|---|---|
| Greenery | 96.692 | 97.657 |
| Water body | 99.442 | 97.094 |
| Barren land | 97.891 | 97.681 |
| Built-up | 98.073 | 99.621 |
| *Macro Avg.* | *98.0246* | *98.013* |

**(a) Landsat-7**

| LC class | Precision | Recall |
|---|---|---|
| Greenery | 97.327 | 97.333 |
| Water body | 99.568 | 97.643 |
| Barren land | 97.387 | 99.419 |
| Built-up | 98.135 | 99.419 |
| *Macro Avg.* | *98.104* | *98.096* |

**(b) Landsat-8**

| LC class | Precision | Recall |
|---|---|---|
| Greenery | 97.8 | 97.558 |
| Water body | 99.541 | 98.088 |
| Barren land | 97.577 | 97.624 |
| Built-up | 97.876 | 99.499 |
| *Macro Avg.* | *98.198* | *98.192* |

**(c) Sentinel-2**

**Table (5) 5-fold cross validation accuracy for RF Classifier trained on IndiaSat.**

| LC class | Precision | Recall |
|---|---|---|
| Greenery | 63.877 | 93.469 |
| Water body | 94.973 | 99.715 |
| Barren land | 86.869 | 20.673 |
| Built-up | 73.523 | 83.582 |
| *Macro Avg.* | *79.810* | *74.360* |

**(a) Bangalore**

| LC class | Precision | Recall |
|---|---|---|
| Greenery | 72.300 | 90.058 |
| Water body | 88.559 | 97.664 |
| Barren land | 91.892 | 34.343 |
| Built-up | 71.470 | 92.884 |
| *Macro Avg.* | *81.055* | *78.737* |

**(b) Delhi**

| LC class | Precision | Recall |
|---|---|---|
| Greenery | 68.700 | 95.750 |
| Water body | 92.742 | 99.138 |
| Barren land | 100.00 | 39.130 |
| Built-up | 95.935 | 98.662 |
| *Macro Avg.* | *98.198* | *98.192* |

**(c) Gurgaon**

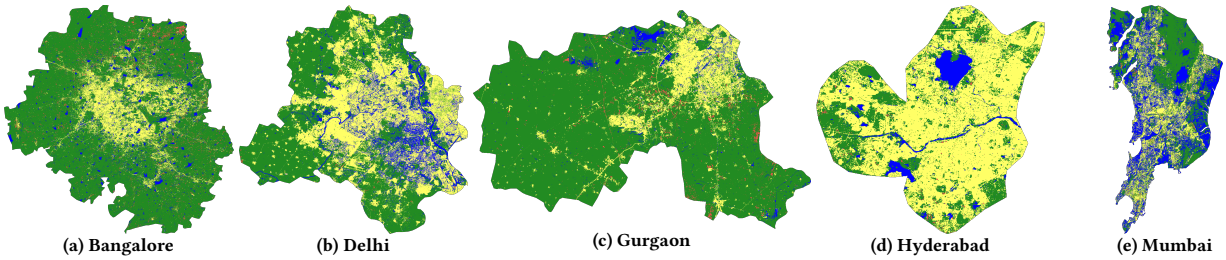
| LC class | Precision | Recall |
|---|---|---|
| Greenery | 71.324 | 95.410 |
| Water body | 94.175 | 100.00 |
| Barren land | 100.00 | 38.500 |
| Built-up | 98.357 | 98.821 |
| *Macro Avg.* | *90.964* | *83.183* |

**(d) Hyderabad**

| LC class | Precision | Recall |
|---|---|---|
| Greenery | 41.402 | 67.760 |
| Water body | 78.676 | 99.442 |
| Barren land | 79.375 | 37.463 |
| Built-up | 88.043 | 61.738 |
| *Macro Avg.* | *71.874* | *66.601* |

**(e) Mumbai**

**Table (6) 4-class test accuracy with Sentinel-2 against out-of-sample ground truth marked for the year 2018.**



**(a) Bangalore** **(b) Delhi** **(c) Gurgaon** **(d) Hyderabad** **(e) Mumbai**

**Figure (7) City-wise 4-class prediction maps for the year 2018.**

| | Model 1 | | | Model 2 | | | Model 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Avg Precision | Avg Recall | Accuracy | Avg Precision | Avg Recall | Accuracy | Avg Precision | Avg Recall | Accuracy |
| Bangalore | 82.428 | 76.462 | 81.65% | 88.168 | 84.366 | 86.31% | 88.857 | 88.471 | 89.73% |
| Delhi | 73.689 | 74.997 | 73.97% | 88.611 | 87.909 | 89.24% | 91.230 | 92.396 | 91.92% |
| Gurgaon | 80.695 | 80.698 | 80.67% | 91.779 | 90.567 | 88.32% | 95.686 | 95.452 | 95.48% |
| Hyderabad | 96.544 | 96.496 | 96.67% | 99.206 | 98.909 | 98.61% | 98.970 | 98.366 | 98.74% |
| Mumbai | 92.954 | 93.685 | 93.58% | 86.067 | 87.306 | 85.33% | 87.937 | 88.623 | 88.55% |
| *Average* | *87.432* | *87.723* | *85.308%* | *90.766* | *89.811* | *89.562%* | *92.479* | *92.661* | *92.884%* |

**Table (7) Comparative evaluation against Goldblatt's baseline data and model.**


[3] Chahat Bansal, Arpit Jain, Phaneesh Barwaria, Anuj Choudhary, Anupam Singh, Ayush Gupta, and Aaditeshwar Seth. 2020. Temporal Prediction of Socio-economic Indicators Using Satellite Imagery. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*. 73–81.

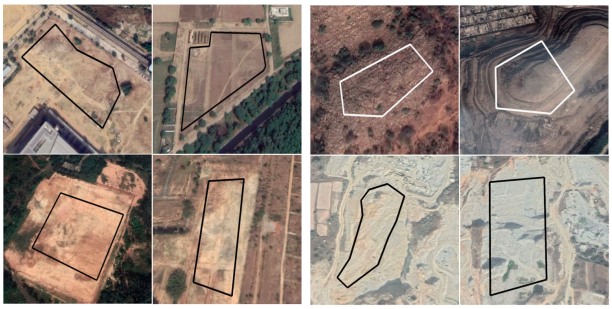
[4] Chahat Bansal, Aditi Singla, Ankit Kumar Singh, Hari Om Ahlawat, Mayank Jain, Prachi Singh, Prashant Kumar, Ritesh Saha, Sakshi Taparia, Shailesh Yadav, et al. 2020. Characterizing the evolution of indian cities using satellite imagery and open street maps. In *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies*. 87–96.

[5] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[6] Chao Chen, Xinyue He, Zhisong Liu, Weiwei Sun, Heng Dong, and Yanli Chu. 2020. Analysis of regional economic development based on land use and land

| | Change Classification | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Landsat-7 | | | Lansat-8 | | | Sentinel-2 | | |
| | Avg Precision | Avg Recall | Accuracy | Avg Precision | Avg Recall | Accuracy | Avg Precision | Avg Recall | Accuracy |
| Bangalore | 74.979 | 66.940 | 73.465% | 75.868 | 63.825 | 61.003% | 84.587 | 80.515 | 79.150% |
| Delhi | 88.844 | 85.378 | 87.315% | 86.289 | 85.391 | 86.381% | 84.542 | 85.944 | 86.225% |
| Gurgaon | 88.145 | 85.137 | 85.408% | 85.561 | 77.454 | 77.906% | 93.995 | 93.775 | 93.899% |
| Hyderabad | 84.875 | 82.694 | 85.831% | 83.882 | 76.357 | 82.546% | 87.275 | 87.241 | 89.117% |
| Mumbai | 76.838 | 56.569 | 73.940% | 80.599 | 73.178 | 81.506% | 88.119 | 70.683 | 81.849% |

**Table (8) Comparing performance of change classification on different cities.**



(a) (b)

**Figure (8) Examples of misclassified barren lands into (a) greenery, and (b) built-up due to their overlapping spectral signatures.**


cover change information derived from Landsat imagery. *Scientific RepoRtS* 10, 1 (2020), 1–16.

[7] J Cihlar. 2000. Land cover mapping of large areas from satellites: status and research priorities. *International journal of remote sensing* 21, 6-7 (2000), 1093–1114.

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3213–3223.

[9] Wakjira Takala Dibaba, Tamene Adugna Demissie, and Konrad Miegel. 2020. Drivers and Implications of Land Use/Land Cover Dynamics in Finchaa Catchment, Northwestern Ethiopia. *Land* 9, 4 (2020), 113.

[10] Laura Dingle Robertson and Douglas J King. 2011. Comparison of pixel-and object-based classification in land cover change mapping. *International Journal of Remote Sensing* 32, 6 (2011), 1505–1529.

[11] Dennis C Duro, Steven E Franklin, and Monique G Dubé. 2012. A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. *Remote sensing of environment* 118 (2012), 259–272.

[12] Maie I El-Gammal, Rafat R Ali, and Rasha Eissa. 2014. Land use assessment of barren areas in Damietta Governorate, Egypt using remote sensing. *Egyptian Journal of Basic and Applied Sciences* 1, 3-4 (2014), 151–160.

[13] Gudina L Feyisa, Henrik Meilby, Rasmus Fensholt, and Simon R Proud. 2014. Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery. *Remote Sensing of Environment* 140 (2014), 23–35.

[14] GEE. [n.d.]. Sentinel-2 MSI: MultiSpectral Instrument, Level-1C. https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2

[15] Ran Goldblatt, Gordon Hanson, Kilian Heilmann, Amit Khandelwal, and Yonatan Vaizman. 2017. Properties of Satellite Imagery for Measuring Economic Activity at Small Geographies. (2017).

[16] Ran Goldblatt, Wei You, Gordon Hanson, and Amit K Khandelwal. 2016. Detecting the boundaries of urban areas in India: A dataset for pixel-based image classification in google earth engine. *Remote Sensing* 8, 8 (2016), 634.

[17] Andrew Head, Mélanie Manguin, Nhat Tran, and Joshua E Blumenstock. 2017. Can Human Development be Measured with Satellite Imagery?. In *ICTD*. 8–1.

[18] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12, 7 (2019), 2217–2226.

[19] Danfeng Hong, Naoto Yokoya, Nan Ge, Jocelyn Chanussot, and Xiao Xiang Zhu. 2019. Learnable manifold alignment (LeMA): A semi-supervised cross-modality learning framework for land cover and land use classification. *ISPRS Journal of Photogrammetry and Remote Sensing* 147 (2019), 193–205.

[20] Juliane Huth, Claudia Kuenzer, Thilo Wehrmann, Steffen Gebhardt, Vo Quoc Tuan, and Stefan Dech. 2012. Land cover and land use classification with TWOPAC: Towards automated processing for pixel-and object-based image classification. *Remote Sensing* 4, 9 (2012), 2530–2553.

[21] Nianxue Luo, Taili Wan, Huaixu Hao, and Qikai Lu. 2019. Fusing high-spatial-resolution remotely sensed imagery and OpenStreetMap data for land cover classification over urban areas. *Remote Sensing* 11, 1 (2019), 88.

[22] Stuart K McFeeters. 1996. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *International journal of remote sensing* 17, 7 (1996), 1425–1432.

[23] Koreen Millard and Murray Richardson. 2015. On the importance of training data sample selection in random forest image classification: A case study in peatland ecosystem mapping. *Remote sensing* 7, 7 (2015), 8489–8515.

[24] PC Pandley, Kiril Manevski, Prashant K Srivastava, and George P Petropoulos. 2018. The use of hyperspectral earth observation data for land use/cover classification: present status, challenges and future outlook. *Hyperspectral Remote Sensing of Vegetation, 1st ed.; Thenkabail, P., Ed* (2018), 147–173.

[25] Suman Patra, Satiprasad Sahoo, Pulak Mishra, and Subhash Chandra Mahapatra. 2018. Impacts of urbanization on land use/cover changes and its probable implications on local climate and groundwater level. *Journal of urban management* 7, 2 (2018), 70–84.

[26] Rutherford V Platt and Lauren Rapoza. 2008. An evaluation of an object-oriented paradigm for land use/land cover classification. *The Professional Geographer* 60, 1 (2008), 87–100.

[27] Jing Qian, Qiming Zhou, and Quan Hou. 2007. Comparison of pixel-based and object-oriented classification methods for extracting built-up areas in arid zone. In *ISPRS workshop on updating Geo-spatial databases with imagery & the 5th ISPRS workshop on DMGISs*, Vol. 36. 163–171.

[28] Parth S Roy, Arijit Roy, Pawan K Joshi, Manish P Kale, Vijay K Srivastava, Sushil K Srivastava, Ravi S Dwevidi, Chitiz Joshi, Mukunda D Behera, Prasanth Meiyappan, et al. 2015. Development of decadal (1985–1995–2005) land use and land cover database for India. *Remote Sensing* 7, 3 (2015), 2401–2430.

[29] ISRO SAC. 2016. Desertification and land degradation Atlas of India (based on IRS AWiFS data of 2011–13 and 2003–05). *Space Applications Centre, ISRO, Ahmedabad, India, Ahmedabad* (2016).

[30] Evelin Priscila Trindade, Marcus Phoebe Farias Hinnig, Eduardo Moreira da Costa, Jamile Sabatini Marques, Rogério Cid Bastos, and Tan Yigitcanlar. 2017. Sustainable development of smart cities: A systematic review of the literature. *Journal of Open Innovation: Technology, Market, and Complexity* 3, 3 (2017), 11.

[31] Compton J Tucker. 1979. Red and photographic infrared linear combinations for monitoring vegetation. *Remote sensing of Environment* 8, 2 (1979), 127–150.

[32] USGS. [n.d.]. Landsat 7 band designations. https://www.usgs.gov/media/images/landsat-7-band-designations

[33] USGS. [n.d.]. Landsat 8 band designations. https://www.usgs.gov/media/images/landsat-8-band-designations

[34] Nisha Varghese and Naveen P Singh. 2016. Linkages between land use changes, desertification and human development in the Thar Desert Region of India. *Land Use Policy* 51 (2016), 18–25.

[35] Cláudia M Viana, Luis Encalada, and Jorge Rocha. 2019. The value of openstreetmap historical contributions as a source of sampling data for multi-temporal land use/cover maps. *ISPRS International Journal of Geo-Information* 8, 3 (2019), 116.

[36] Yichun Xie, Zongyao Sha, and Mei Yu. 2008. Remote sensing imagery in vegetation mapping: a review. *Journal of plant ecology* 1, 1 (2008), 9–23.

[37] Gao Yan, J-F Mas, BHP Maathuis, Zhang Xiangmin, and PM Van Dijk. 2006. Comparison of pixel-based and object-oriented image classification approaches—a case study in a coal fire area, Wuda, Inner Mongolia, China. *International Journal of Remote Sensing* 27, 18 (2006), 4039–4055.