Machine Learning Approaches for Slum Detection Using Very High Resolution Satellite Images

Krishna Karthik Gadiraju
Department of Computer Science
North Carolina State University
Raleigh, USA
kgadira@ncsu.edu

Ranga Raju Vatsavai
Department of Computer Science
North Carolina State University
Raleigh, USA
rrvatsav@ncsu.edu

Nikhil Kaza
Department of City and Regional Planning
The University of North Carolina
Chapel Hill, USA
nkaza@unc.edu

Erik Wibbels
Trinity College of Arts & Sciences
Duke University
Durham, USA
e.wibbels@duke.edu

Anirudh Krishna
Sanford School of Public Policy
Duke University
Durham, USA
ak30@duke.edu

***Abstract*—Detecting informal settlements has become an important area of research in the past decade, owing to the availability of high resolution satellite imagery. Traditional per-pixel based classification methods provide high degree of accuracy in distinguishing primitive instances such as buildings, roads, forests and water. However, these methods fail to capture the complex relationships between neighboring pixels that is necessary for distinguishing complex objects such as informal and formal settlements. In this paper, we perform several experiments to compare and contrast how various per-pixel based classification methods, when combined with various features perform in detecting slums. In addition, we also explored a deep neural network, which showed better accuracy than the pixel based methods.**

***Index Terms*—remote-sensing, image-classification, informal-settlements**

I. Introduction

Detecting and studying different types of urban settlements is an active area of research. Several factors have contributed to the interest in this area of research. Firstly, there has been a rapid rise in the availability of very high resolution (VHR) imagery. The fine spatial resolution of such VHR imagery has enabled research in identifying complex patterns such as commercial complexes, informal settlements and formal settlements. Secondly, as described in [1], it is projected that by the year 2050, two out of three people will live in a city. As described in [2], this rapid migration towards the cities due to higher quality of life in developing countries leads to adverse impacts on climate, environment and life, due to higher carbon emissions, traffic congestion, agricultural and forest destruction, etc. As a result, there is a great need to detect and study the different types of settlements, and in particular, the informal settlements. Informal settlements (also known as slums or barrios or shantytowns or low income settlements) refer to unplanned, unauthorized and/or unstructured homes [3]. According to [4], in 2014, an estimated 880 million urban residents lived in slums, compared to 792 million in 2000. These numbers are even more important from the perspective of a developing nation such as India, where nearly 22% of the population lives in these information settlements [5]. In this paper, we demonstrate methods to detect different types of informal settlements in and around the city of Bengaluru, Karnataka, India. In the following sections, we first discuss related work in the area of informal neighborhood classification. We then discuss and evaluate various machine learning approaches. Finally, we summarize our results and provide directions for future research. In the rest of this paper, we refer to the low-income settlements as informal settlements or slums interchangeably.

II. Related Work

Traditional classification schemes work with pixels (aka single instances). They are good at identifying thematic classes (such as urban, forests, crops, etc., or subclasses such as high/low density urban, hardwood forest, conifer forest, etc.). Several works such as [6] exist in literature that deal with classification of such thematic classes using single instance learners. However, urban neighborhood classification requires analyzing image patch (group of pixels) as a unit in order to model the spatial context. There are two distinct approaches to model spatial context: one is through extracting features that capture spatial contextual properties and use traditional classification schemes (single instance learners), the other is to use classification schemes (e.g., convolutional neural networks [7] or multiple instance learning [3], [8]) that work with image patches (as compared to pixels). In this work, we analyzed both approaches and compared various state of the art methods in order to characterize their performance in slum identification. In addition to classification, we also studied temporal dynamics in slums. Due to limited availability of VHR imagery, our study focused on identifying changes by analyzing best available imagery (Landsat 7, 15 meter spatial resolution, year 2002).

2375-9259/18/$31.00 ©2018 IEEE
DOI 10.1109/ICDMW.2018.00198

As described in [3], the primary motivation behind considering patch-based approaches in comparison to pixel-based approach for classes such as informal and formal neighborhoods is that traditional pixel based learners fail to capture the complex spatial relationships that are associated with the aforementioned classes. In order to identify these classes, a bigger region (or a patch) than a single pixel needs to be considered in order to capture the complex properties associated with data of these classes such as: densely packed buildings, lack of vegetation and roads, etc.

In this paper, we demonstrate both pixel-based and patch-based CNN approaches for detecting different types of low-income settlements in VHR imagery. While [7] also performs detection of informal settlements in VHR imagery using a convolutional neural network (patch-based approach), our work is different from [7] in that we are considering a smaller patch size, and are also performing classification of a larger variety of informal settlements. In addition, we perform some additional analyses to answer specific questions about the informal settlements which is not performed in [7].

III. Types of Classification

In this section, we describe the different types and levels of classification performed in this paper. In addition to the distinction of pixel-based vs. patch-based classification, we also make another distinction in the type of classification under which both the pixel and patch based classifications performed. We define two types of classification:

1) Multi-Class Classification*: in this classification, we progressively increase the number of classes in each level:
   - Level 1: data of type Urban vs Other
   - Level 2: data of type Formal vs Informal vs Other
   - Level 3: data of type Single Story vs Multi-Story vs Semi-Permanent vs Temporary vs Formal vs Other
2) Hierarchical Classification*: In multi-class classification, we observed that the confusion between classes in each level increases as the level (number of classes) increases. To overcome this limitation, we devised a hierarchical classification, in which finer classes were confined to coarser class regions (masks) from previous levels.

*To distinguish between multi-class and hierarchical classifications, we are using MC-level [1-3] for multi-class classification scheme and HC-level [1-3] respectively to denote classification scheme.

IV. Feature Generation

In order to perform both pixel-based and patch based image analysis, the following features were generated:

- Haralick Texture Features: It has been well documented in literature that Haralick texture features [9] are highly efficient in distinguishing urban from other classes such as vegetation and water.
- NDBI (Normalized Difference Built-Up Index) [10]: This is measured using the formula:

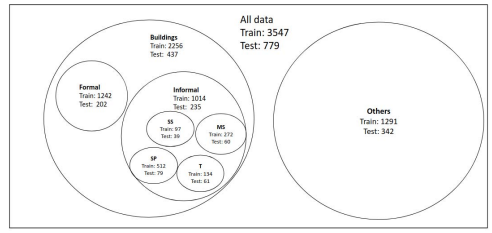


Fig. 1. Division of data collected into various classes

$NDBI = \frac{SWIR-NIR}{SWIR+NIR}$, where SWIR and NIR refers to the Short Wave Infra Red and Near Infra Red bands of the VHR image. This feature also helps distinguish between urban class and other classes such as vegetation and water.
- Edge Density: Edge density is calculated by first detecting edges in the entire image using the Canny Edge Detection [11] method. Then the average of all edges within a neighborhood of a pixel is assigned to the pixel as its edge density.
- Pansharpened Bands using 2m Multi-Spectral Image + 0.5m panchromatic image: The 8 Multi-Spectral bands (from the 2m VHR image), which are originally of 2m resolution are combined with the 0.5 m panchromatic band (from the same VHR image) using the RCS (Ratio Component Substitution) method available in the Orfeo Toolbox (OTB). The advantage of using this method is increased spatial information available to due to the fact that the bands will be of much finer resolution.

V. Experiments

The class labels used in our experiments are defined as shown below:

- Buildings/Urban (U) - refers to all types of built up area
- Formal (F) - refers to formal type constructions
- Informal (IF) - refers to slums/shanty towns
- Single-Story (SS) - refers to informal constructions of type single-story
- Multi-Story (MS) - refers to informal constructions of type multi-story
- Semi-Permanent (SP) - refers to informal constructions of type semi-permanent
- Temporary (T) - refers to temporary informal settlements
- Background/Others (O) - data that doesnt belong to the aforementioned classes, such as vegetation, open land, water etc. is collectively stored as Other.

A. Multi-Class Classification

In multi-class classification we perform the following classifications:

- MC-Level 1: Urban (U) vs Other (O)
- MC-Level 2: Formal (F) vs Informal (IF) vs Other (O)
- MC-Level 3: Single Story (SS) vs Multi-Story(MS) vs Semi-Permanent(SP) vs Temporary(T) vs Formal(F) vs Other (O)

1) Pixel Based Classifiers: For pixel-based or single-instance classifiers, the following classifiers: Naive Bayes (NB), Decision Tree (DT), K-Nearest Neighbours (KNN), Multi Layer Perceptron (MLP), Gradient Boosting (XGB), Random Forest (RF) and Adaboost Classifier (ADB) were used.

The entire data collected was divided into training and testing as shown in Figure 1. Within the training data, grid-search was performed using 10-fold cross validation and negative log loss was used as the metric to find the optimal hyperparameters for classifiers such as XGB, RF, MLP and KNN. Classification was performed on the test data and accuracy measures, including overall accuracy, precision, recall and f-measure were noted for all the three levels of classification.

TABLE I
MC-LEVEL 1 CLASSIFICATION OUTCOMES

Classifier	*Overall Accuracy*	*Precision*	*Recall*	*F-Measure*
NB	0.96	0.96	0.96	0.96
DT	0.98	0.98	0.98	0.98
KNN	0.97	0.97	0.97	0.97
MLP	0.98	0.98	0.98	0.98
XGB	0.99	0.99	0.99	0.99
ADB	0.99	0.99	0.99	0.99
RF	0.98	0.98	0.98	0.98
CNN	1	1	1	1

TABLE II
MC-LEVEL 2 CLASSIFICATION OUTCOMES

Classifier	*Overall Accuracy*	*Precision*	*Recall*	*F-Measure*
NB	0.77	0.78	0.77	0.76
DT	0.80	0.81	0.80	0.80
KNN	0.74	0.73	0.74	0.73
MLP	0.77	0.80	0.77	0.75
XGB	0.83	0.84	0.83	0.83
ADB	0.77	0.77	0.77	0.77
RF	0.83	0.84	0.83	0.83
CNN	0.86	0.87	0.86	0.86

TABLE III
MC-LEVEL 3 CLASSIFICATION OUTCOMES

Classifier	*Overall Accuracy*	*Precision*	*Recall*	*F-Measure*
NB	0.61	0.69	0.61	0.62
DT	0.69	0.69	0.69	0.68
KNN	0.66	0.62	0.66	0.61
MLP	0.69	0.54	0.68	0.59
XGB	0.75	0.74	0.75	0.72
ADB	0.66	0.55	0.66	0.58
RF	0.74	0.72	0.74	0.71
CNN	0.71	0.71	0.71	0.70

2) Patch Based Classifiers: In contrast to pixel-based classification methods, where a label is assigned to each pixel, a patch-based classification technique assigns a label to a group of pixels. By considering groups of pixels as a single entity, complex interactions between individual objects in a patch are well captured by the patch-based classifiers. In this work, the entire 0.5m mosaic for the Bengaluru region is sub-divided into grids or patches of size 40 * 40 pixels, and a label is assigned to each patch. This patch size is experimentally determined based on typical sizes of structures and the size of the ground truth polygons. A convolutional neural network (CNN) was designed to perform Levels 1,2 and 3 classification. Figure 2 depicts the general structure of the CNN used for Levels 1,2 and 3 classification. While all the three networks have a similar structure with the following properties:

- The input image is fed to 3 2D Convolutional layers, followed by a dropout layer, which is further followed by 3 2D convolution layers followed by a max pooling layer, which is then followed by 1 more 2D convolution layer. These layers are then followed by a max pool layer, followed by a fully connected layer, which is then followed by a dropout layer, which is then followed by two fully connected layers. The dropout layers are included to ensure that the CNN doesnt overfit on the training data (regularization). The maxpool layers are included to reduce the computational complexity.
- Each of the 2D convolutional layers is regularized using L2 regularization, and each of them is initialized using Xavier initialization. RELU activation is used for each of the 2D convolutional layers and the first two fully connected layers, while the final fully connected layer uses softmax activation to get the predicted probabilities for each class.
- Categorical cross entropy is used as the loss function and the Adam optimizer is used for optimizing the CNN.

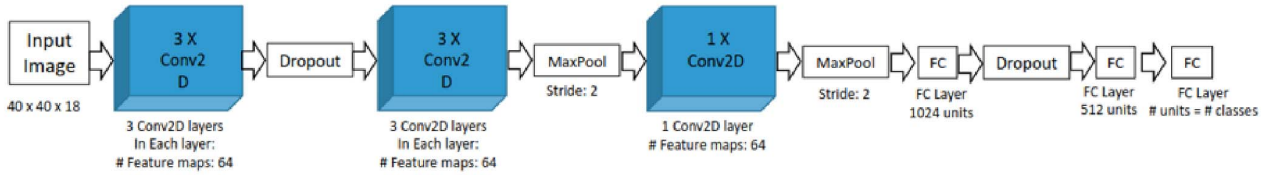
The networks used for Levels 1, 2 and 3 classification are different in the fact that different filter sizes are used for each of the levels of classification. The filter sizes were chosen based on optimal training/test accuracy by trial and error. While Level 1 classification had a filter size of 7 for the first 5 2D convolutional layers and size 5 for the last two, Levels 2 and 3 had a filter size of 7 for all the 2D convolutional layers. 30% of the training data is used for validation, and the CNNs are trained for 256 epochs. Data augmentation is performed by randomly rotating the training images by 90 degrees and also randomly flipping the training images left to right and top to bottom.

B. Hierarchical Classification

From the confusion matrices shown above, he following observations are clearly evident:

- There is misclassification/confusion between urban and other classes which are getting carried over from MC-Level 1 to MC-Level 2 and 3 classifications.
- There is misclassification/confusion between different types of informal classes such as single-story and multi-story and formal class, which are getting carried over from MC-Level 2 to MC-Level 3 classification.

In other words, the misclassification which is appearing at MC-Level 1 is being carried over to MC-Level 2, and the misclassification appearing at MC-Level 2 is being carried over to MC-Level 3. In order to overcome this confusion (or



-2

Fig. 2. Structure of CNN used in this work

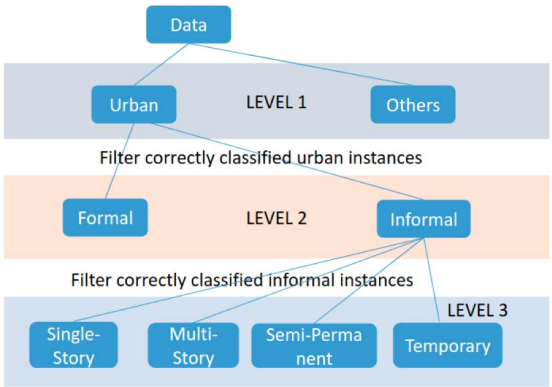


Fig. 3. Levels of Hierarchy

repeat misclassifications at each classification), we perform classification in a hierarchical manner, by first performing classification between Urban and Other classes. We call this HC-Level 1 classification. We then filter the original training and test data to identify the correctly classified Urban instances only and then perform classification on Informal vs Formal data using this filtered data from HC-Level 1 classification. We call this HC-Level 2 classification. We then filter the training and test data from this step and then perform classification on Single-Story vs Multi-Story vs Semi-Permanent vs Temporary classes using this newly filtered data. We call this HC-Level 3 classification. Figure 3 depicts this hierarchical classification at the three different levels. Using the same classifiers as mentioned in V-A1, we perform training and testing at the three levels: HC-Level 1, 2 and 3. The results are described in tables shown below.

TABLE IV
HC-LEVEL 1 CLASSIFICATION OUTCOMES

Classifier	*Overall Accuracy*	*Precision*	*Recall*	*F-Measure*
NB	0.96	0.96	0.96	0.96
DT	0.98	0.98	0.98	0.98
KNN	0.97	0.97	0.97	0.97
MLP	0.98	0.98	0.98	0.98
XGB	0.99	0.99	0.99	0.99
ADB	0.99	0.99	0.99	0.99
RF	0.99	0.99	0.99	0.99
CNN	1	1	1	1

C. Change Detection

In this section, the primary objective is to identify the status of a current informal region (year 2016) in the year 2002.

TABLE V
HC-LEVEL 2 CLASSIFICATION OUTCOMES

Classifier	*Overall Accuracy*	*Precision*	*Recall*	*F-Measure*
NB	0.62	0.63	0.62	0.58
DT	0.63	0.63	0.63	0.63
KNN	0.57	0.56	0.57	0.56
MLP	0.59	0.59	0.59	0.59
XGB	0.70	0.70	0.70	0.70
ADB	0.69	0.69	0.69	0.69
RF	0.72	0.72	0.72	0.72
CNN	0.70	0.70	0.70	0.70

TABLE VI
HC-LEVEL 3 CLASSIFICATION OUTCOMES

Classifier	*Overall Accuracy*	*Precision*	*Recall*	*F-Measure*
NB	0.45	0.46	0.45	0.38
DT	0.36	0.36	0.36	0.36
KNN	0.36	0.34	0.36	0.32
MLP	0.42	0.34	0.42	0.35
XGB	0.44	0.44	0.46	0.45
ADB	0.48	0.34	0.42	0.35
RF	0.50	0.49	0.50	0.47
CNN	0.57	0.59	0.57	0.52

Based on the status of the 2002 image category (class), we grouped the status of the informal settlements into:

- New informal settlements: informal settlements that were constructed after 2002.
- long-existing informal settlements: informal settlements that existed both in 2002 and 2016.

For the sake of brevity, New informal settlements will be identified as NIF and long-existing informal settlements will be identified as OIF henceforth. In order to identify informal neighborhoods of type NIF and OIF, we perform the following steps:

- **Step 1:** identify built-up and non-built up area in the Landsat 7 data from the year 2002.
- **Step 2:** identify informal neighborhoods in the VHR image from the year 2016. (as described in the previous sections)
- **Step 3:** Compare the outcomes from steps 1 and 2 to identify NIF and OIF regions

The LANDSAT-7 images for the study region are collected for February 2002. The 6 bands of LANDSAT-7, which are of 30m resolution are pansharpened to 15m resolution using the panchromatic band.

1) Step 1: Identify built-up and non-built up area in LANDSAT-7 data from 2002: We extract a building mask from the LANDSAT-7 data to differentiate built-up area from

other types such as open land or barren land or vegetation. Given that the spectral signature of a building is very close when compared to that of a barren land, it is necessary that we generate more features in addition to the existing spectral features. The following additional features are generated:

- Haralick Texture Features [9]: In addition to the simple haralick textures described in IV, advanced Features [Mean, Variance, Dissimilarity, Sum Average, Sum Variance, Sum Entropy, Difference of Entropies, Difference of Variances, IC1 and IC2] and higher Features [Short Run Emphasis, Long Run Emphasis, Grey-Level Nonuniformity, Run Length Nonuniformity, Run Percentage, Low Grey-Level Run Emphasis, High Grey-Level Run Emphasis, Short Run Low Grey-Level Emphasis, Short Run High Grey-Level Emphasis, Long Run Low Grey-Level Emphasis and Long Run High Grey-Level Emphasis] are also generated for the LANDSAT 7 image.
- NDBI, as described in IV.
- Pansharpened bands using the pansharpened LANDSAT 7 band, which makes the rest of the bands 15m in resolution.

In order to create a building mask using the LANDSAT-7 image, we manually select around 2845 data points and divide them into training and testing. Table VII depicts the train-test data split per class.

TABLE VII
LANDSAT-7 TRAIN-TEST SPLIT

Class	*# Training*	*# Test*
Buildings	1196	242
Background/Other	1125	282

We use the same classifiers as described in V-A1. Finally, we use a majority vote classifier, which would assign the label to a data point based on what the majority of the aforementioned classifiers classify it as. The primary reason behind using a majority vote classifier is to ensure to capture the efficiency of all the classifiers In other words, certain classifiers may classify a section of data better than the others; by combining all the classifiers, we can ensure that the efficiency of each classifier is captured effectively. Table VIII displays the accuracy measures (overall accuracy, precision, recall and F-Measure) for each of the classifier for the test data for the pixel-based classification on the LANDSAT-7 image.

TABLE VIII
LANDSAT-7 CLASSIFICATION OUTCOMES

Classifier	*Overall Accuracy*	*Precision*	*Recall*	*F-Measure*
NB	0.97	0.97	0.97	0.97
DT	0.97	0.97	0.97	0.97
KNN	0.97	0.97	0.97	0.97
MLP	0.99	0.99	0.99	0.99
XGB	0.99	0.99	0.99	0.99
ADB	0.99	0.99	0.99	0.99
RF	0.99	0.99	0.99	0.99
Majority Vote	1	1	1	1

2) Step 2: Extraction of Informal mask from VHR: is as described in V-A and V-B.

3) Step 3: Identify settlements of type NIF and OIF: In order to identify the NIF and OIF regions, we first overlay the VHR and LANDSAT 7 images over each other and then:

- New Informal Settlements: In order to identify NIF settlements, we search for those informal neighborhoods in the VHR at whose location, the LANDSAT 7 image was classified as open land.
- Long-Existing Informal Settlements: We achieve this by following a procedure similar to the one followed for identifying NIF settlements. The only difference is that, we only look for those informal regions in 2012 which were identified as built-up areas in 2002. Given the low resolution of the LANDSAT-7 imagery, it will not be possible to identify if these identified regions were actually of type informal in 2002. In order to filter the OIF settlements from the identified regions, we perform the following tasks:
  1) draw polygons around the regions identified by this experiment.
  2) use the historical imagery (which has a higher resolution than LANDSAT-7) available nearest to the years 2016 and 2002 available in Google Earth, visually verify if the polygons generated in (1) were actually *long-existing informal settlements (OIF)* or *formal settlements misclassified as informal settlements (MIF)*

The outcomes of these experiments are described in the next section.

D. Results and Analysis

- Analysis in terms of classification accuracy:
  - MC-Level 1 Classification: Then CNN achieved approximately 99.74% accuracy when separating urban class from all other classes, which is marginally better than the best single instance method, which was the XGB classifier and the ADB classifier.
  - MC-Level 2 Classification: The CNN achieved approximately 86% accuracy when separating complex labels such as Formal and Informal neighborhoods. From the confusion matrices in Tables IX, X, it is clear that the CNN achieves a better classification compared to the best pixel-based classifier, which is the XGB Classifier. The F1-score for informal class using the CNN is better than the F1-score achieved for the XGB Classifier. While the precision of the XGB classifier for the informal neighborhoods is higher, it has a much lower recall compared to the CNN, and as a result, the F1 score of the CNN is better.

TABLE IX
CONFUSION MATRIX FOR MC-LEVEL 2 CLASSIFICATION USING CNN

Actual/Predicted	*Informal*	*Formal*	*Other*
Informal	199	34	2
Formal	70	132	0
Other	0	0	342

TABLE X
CONFUSION MATRIX FOR MC-LEVEL 2 CLASSIFICATION USING XGB

Actual/Predicted	*Informal*	*Formal*	*Other*
Informal	148	86	1
Formal	38	164	0
Other	0	2	340

- MC-Level 3 Classification: The CNN achieved approximately 71.2% accuracy, while the best single-instance classifier, XGB achieved approximately 74% overall accuracy. However, as seen in Tables XI and XII it is to be noticed that Single Story informal neighborhoods (SS) are better identified in the CNN, while Temporary informal neighborhoods (T) is better in the XGB Classifier. While in CNN, more Single-Story instances are being misclassified as Semi-Permanent, in the XGB classifier, they are being misclassified as Formal neighborhoods. The justification for misclassification between single story and semi-permanent informal settlement types by the CNN can be justified as shown below:
  1) Consider Figure 4 (a) where the region covered by a yellow rectangle refers to a Semi-Permanent neighborhood in ground-truth, while the region covered by a red rectangle refers to a Single-Story neighborhood in the ground truth. Visually, both these patches look very similar. As a result, the Single-Story patch is misclassified as a Semi-Permanent patch. From Figure 4 (a), it is evident how all the bands of the VHR image overlap when comparing the data instances from single-story and semi-permanent.
  2) In Figure 4 (b), the blue rectangles on the left represent ground truth data for Multi-Story informal settlements, while on the image on right, the green rectangles represent ground truth for formal neighborhood. It can be observed that ground truth for both these types of patches look very similar. From Figure 4 (b), it is very evident how the bands overlap for the data instances of type multi-story and formal.
  3) The low accuracy of the CNN at MC-Level 3 classification can also be attributed to the fact there is very little amount of training data for certain classes such as Temporary, for which, we can see that the method is not performing accurately.
- Comparison between Multi-Class and Hierarchical approaches: Consider the confusion matrices for multi-class and hierarchical approaches from Tables IX- XVI. Given that the training and test datasets will be different in both the approaches (since the data is filtered in the hierarchical approach), a direct comparison in terms of accuracy is not possible. However, a look at the confusion matrices reveals the advantage of using the hierarchical approach, in the fact that greater number of instances of classes SS, MS, SP and T are classified correctly, when compared to the multi-class approach. This is because, in the multi-class approach, a majority instances of classes SS, MS, SP and T are classified as type F. However, since we are filtering out misclassified instances of type IF and also instances of type F in the earlier step (HC-Level 2 classification), these errors are minimized. As a result, we achieved a better classification. From Table VI, we can also observe that as the complexity of the labels increases in HC-Level 3 classification, the CNN (patch-based approach) performs much better than the pixel-based approaches.
- Change Detection: From Table VIII, we can notice that amongst the individual classifiers, the gradient-boosting based XGB classifier performs the best, achieving a 99.23% accuracy. When the outcomes of all the classifiers are combined together with a majority vote, we notice that we achieve the highest accuracy, with 99.6%. Given that using the majority vote gave the best accuracy, we use this classifier to generate the building mask over the entire image. We then proceed to identifying NIF and OIF settlements. We identify NIF and OIF settlements as described in V-C.
  - NIF Settlements: Figures 5 (a,b) depict a few examples of NIF neighborhoods. Google Earth images from previous years and 2016 have been provided to give the reader a visual reference of how data has changed over the time period.
  - OIF and MIF Settlements: Figures 5 (c,d) depict a few examples of type OIF while Figures 5 (e, f) depict examples of type MIF.

TABLE XI
CONFUSION MATRIX FOR MC-LEVEL 3 CLASSIFICATION USING CNN

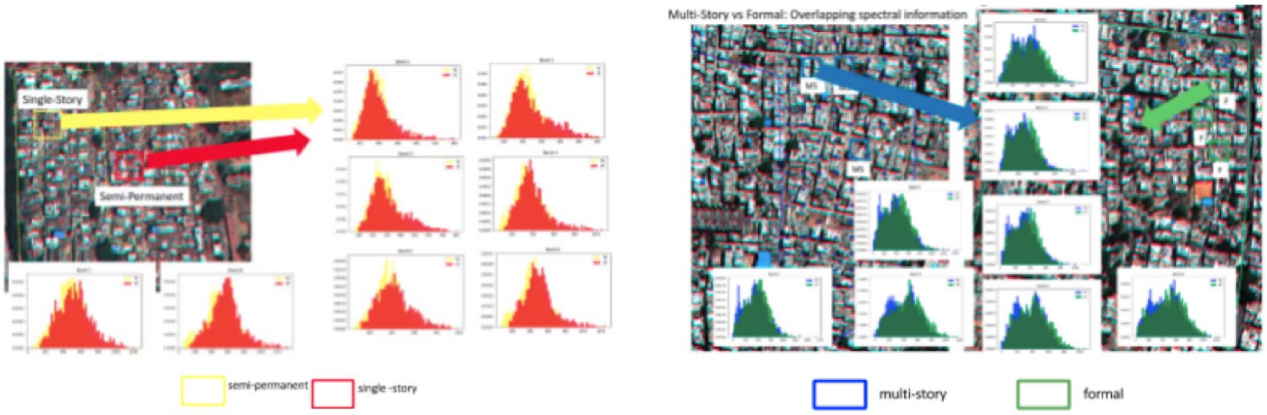
Actual/Predicted	*SS*	*MS*	*SP*	*T*	*F*	*O*
SS	11	2	17	3	6	0
MS	3	16	5	2	32	2
SP	5	2	44	5	23	0
T	5	4	39	6	3	0
F	2	31	24	4	136	5
O	0	0	0	0	6	342

TABLE XII
CONFUSION MATRIX FOR MC-LEVEL 3 CLASSIFICATION USING XGB

Actual/Predicted	*SS*	*MS*	*SP*	*T*	*F*	*O*
SS	1	3	8	4	23	0
MS	0	22	10	1	27	0
SP	2	0	39	2	36	0
T	0	0	23	11	22	1
F	0	9	17	1	174	1
O	0	0	1	0	6	335

TABLE XIII
CONFUSION MATRIX FOR HC-LEVEL 2 CLASSIFICATION USING CNN

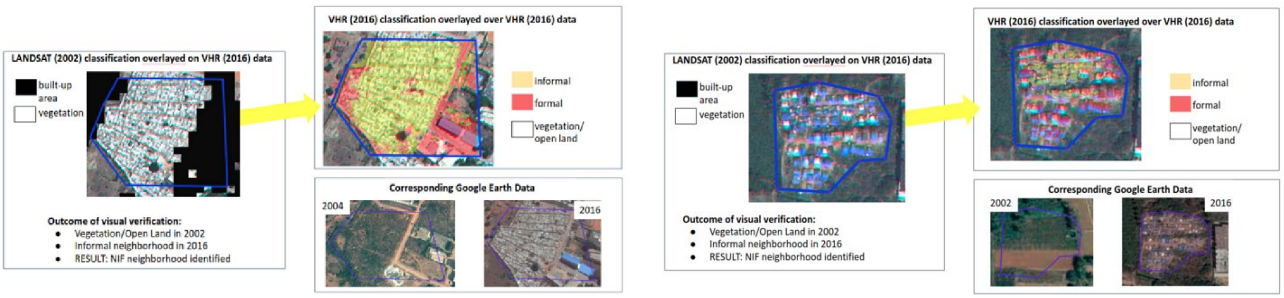
Actual/Predicted	*Informal*	*Formal*
Informal	229	151
Formal	92	336



(a) Semi Permanent vs Single Story

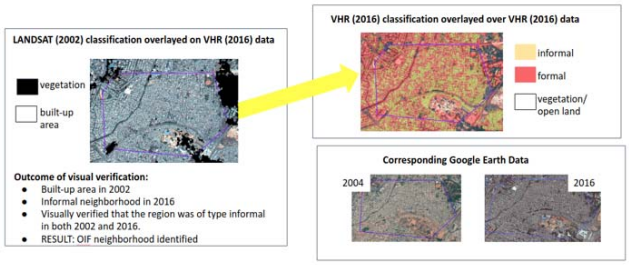
(b) Multi Story vs Formal

Fig. 4. Comparison of spectral signatures of different bands for different classes

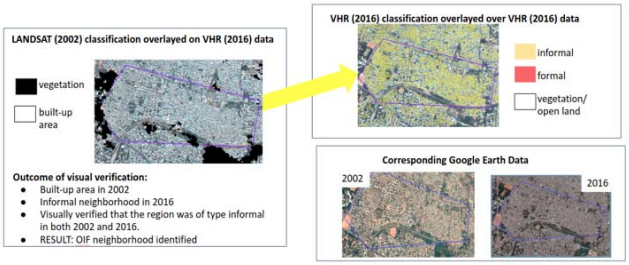


(a) NIF Neighborhood

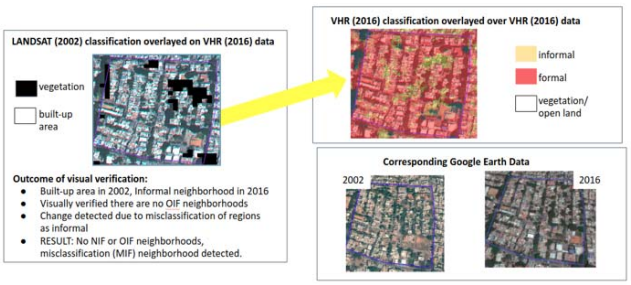
(b) NIF Neighborhood



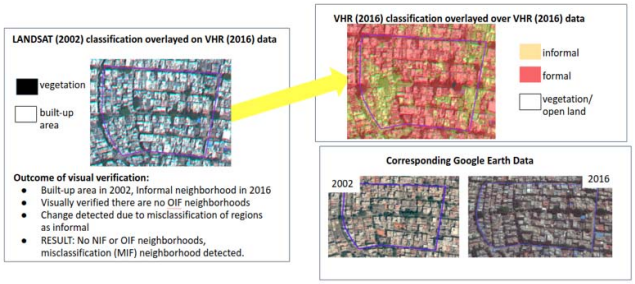
(c) OIF Neighborhood



(d) OIF Neighborhood



(e) MIF Neighborhood



(f) MIF Neighborhood

Fig. 5. Example NIF, OIF and MIF neighborhoods detected using our method

TABLE XIV
CONFUSION MATRIX FOR HC-LEVEL 2 CLASSIFICATION USING XGB

Actual/Predicted	*Informal*	*Formal*
Informal	238	119
Formal	115	336

TABLE XV
CONFUSION MATRIX FOR HC-LEVEL 3 CLASSIFICATION USING CNN

Actual/Predicted	*SS*	*MS*	*SP*	*T*
SS	21	5	9	1
MS	13	29	43	3
SP	1	2	87	1
T	1	3	22	2

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we have studied the performance of several well known machine learning approaches and compared pixel-based and patch-based approaches to detecting different types of informal settlements in VHR imagery. We also demonstrated a hierachical approach to classifying different types of settlements, in which the CNN (patch based approach) showed better results than the pixel based approaches. The research conducted in this paper can be further extended by:

- Improving the accuracy of all the classifiers by adding new features: As shown in the previous sections, since some of the data of class formal and class informal visually looks very similar to each other, in addition to the features generated such as edge density, NDBI and Haralick texture features, there is a need to generate additional unique features that can better distinguish between the two classes. Some possible additional features include ones such as night time lighting information for the data [12]. Since formal neighborhoods are more likely to have proper lighting system when compared to informal neighborhoods, collecting this data may improve the accuracy of the classifiers.
- Improving accuracy of the CNN: Research can be conducted into improving the accuracy of the CNN in the following ways:
  - by data augmentation: In addition to the data augmentation techniques being used currently, additional data augmentation techniques such as randomly resampling the data for classes with minimum number of instances such as Single Story and Temporary may improve the classification accuracy of the CNN.
  - by using better loss functions: develop a custom loss function instead of the categorical cross entropy function which will penalize misclassification of classes that are important to us (such as informal instances of type temporary)
  - use larger patch sizes: using a patch size larger than 40 * 40 pixels would enable us to capture greater neighborhood information, as well as use deeper networks. However, ground truth needs to be updated to reflect the new patch size.
  - transfer learning: the hierarchical classification approach described in the previous section can be modified to a transfer learning approach, where a model pre-trained on Level 1 classification is used for training Level 2 and this model is used for training Level 3.

TABLE XVI
CONFUSION MATRIX FOR HC-LEVEL 3 CLASSIFICATION USING RF

Actual/Predicted	*SS*	*MS*	*SP*	*T*
SS	34	26	54	1
MS	18	37	16	0
SP	17	19	91	6
T	8	0	29	7

ACKNOWLEDGMENT

This research is sponsored by the Omidyar Network. We would like to thank DigitalGlobe for imagery grant and graduate students at UNC and Duke for various preprocessing tasks and interpretation of results. We would like to thank all field researchers for collecting and processing ground truth data. We would also like thank Dr. Frank Mueller from NCSU for granting us access to NCSU's ARC cluster to perform our experiments.

REFERENCES

[1] D. Roy and D. Bernal, "An exploratory factor analysis model for slum severity index in Mexico City," Tech. Rep., 2018.
[2] B. Pradhan, "Spatial Modeling and Assessment of Urban Form Analysis of Urban Growth: From Sprawl to Compact Using Geospatial Data," Tech. Rep.
[3] R. R. Vatsavai, "Gaussian multiple instance learning approach for mapping the slums of the world using very high resolution imagery," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13*, 2013.
[4] U. N. publication issued by the DeSA, "The Sustainable Development Goal Report 2017," Tech. Rep.
[5] O. Kit and M. Lüdeke, "Automated detection of slum area change in hyderabad, india using multitemporal satellite imagery," *ISPRS journal of photogrammetry and remote sensing*, vol. 83, pp. 130–137, 2013.
[6] A. Sekertekin, A. Marangoz, and H. Akcin, "Pixel-based classification analysis of land use land cover using sentinel-2 and landsat-8 data," *Forest*, vol. 80, no. 78.13, pp. 82–71, 2017.
[7] N. Mboga, C. Persello, J. Bergado, and A. Stein, "Detection of Informal Settlements from VHR Images Using Convolutional Neural Networks," *Remote Sensing*, 2017.
[8] R. R. Vatsavai, "Scalable multi-instance learning approach for mapping the slums of the world," in *High Performance Computing, Networking, Storage and Analysis (SCC), 2012 SC Companion:*. IEEE, 2012, pp. 833–837.
[9] R. M. Haralick, K. Shanmugam *et al.*, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.
[10] Y. Zha, J. Gao, and S. Ni, "International Journal of Remote Sensing Use of normalized difference built-up index in automatically mapping urban areas from TM imagery Use of normalized difference built-up index in automatically mapping urban areas from TM imagery," 2010. [Online]. Available: http://www.tandfonline.com/action/journalInformation?journalCode=tres20
[11] J. Canny, "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1986.
[12] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," Tech. Rep. [Online]. Available: http://science.sciencemag.org/