Identifying and Classifying Informal Settlements in Argentina

In the past few decades urbanization has been bringing a large number of new residents to cities. This rapid urban expansion caused the inadequate developments of infrastructure, such as affordable housing, public transport and utilities.

Informal settlements known as slums emerged which had inadequate infrastructure and insufficient living areas inhabited by low-income urban residents.

The UN states that today around one-quarter of the world's population live in slums.

Motivation

Slum identification within a given city will help the authorities in overall city planning efforts and also to provide management accordingly.

Objective

In this project, my goal is to identify the slum areas in the city of Buenos Aires, Argentina through remote sensing with machine learning algorithms.

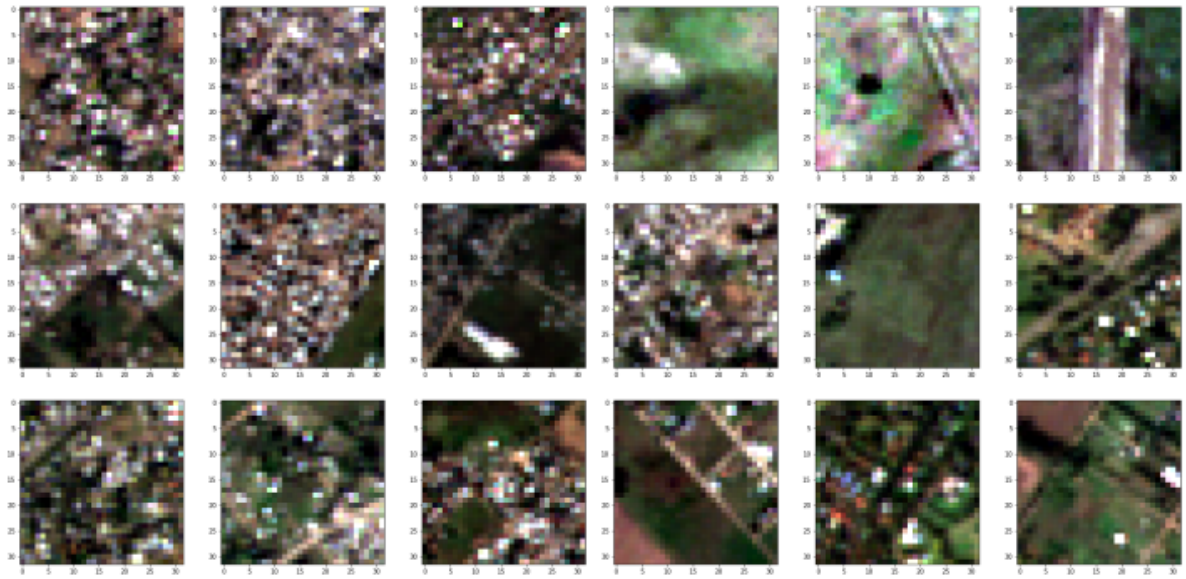
Process

I will first use satellite images from Buenos Aires district to train a machine learning algorithm that identifies which regions are slums, then I will use the trained algorithm to detect slums of Buenos Aires over the span of the past few years.

Data

Training Data

The training data comes from Kaggle Competition: Slums and informal settlements detection https://www.kaggle.com/datasets/fedebayle/slums-argentina . The dataset contains georeferenced images about urban slums and informal settlements for two districts in Argentina: Buenos Aires and Córdoba.



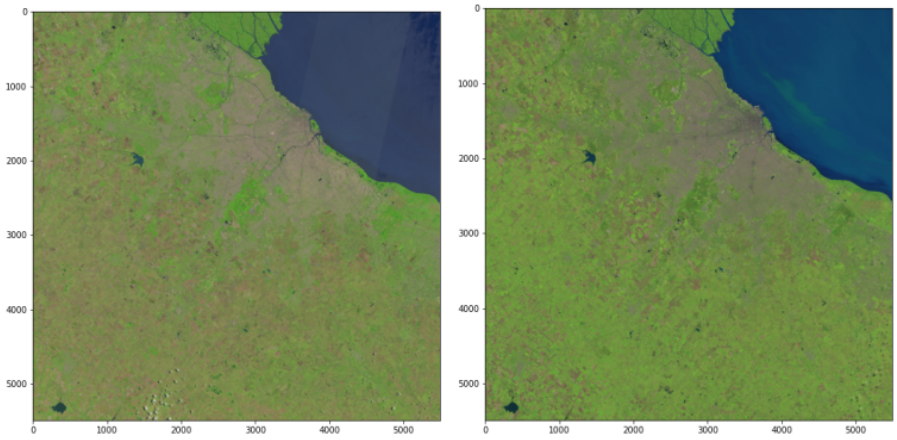
The image of Cordoba was taken on 2017-06-09 (37 out of the 13,3014 images are labeled as slums) and the images of Buenos Aires on 2017-05-04, (1008 out of 46,047 images are labeled as slums).
This dataset is highly unbalanced.

Each image is from Sentinel 2 with 32 by 32 pixels containing 4 bands (bands 2, 3, 4, 8A) at a 10-meter resolution. Images are in .tif format.

Application Data

I applied my model to track how the slums and settlements changed in Buenos Aires over the years. I downloaded two Sentinel-2 images of the region taken on Dec 20th 2017 and April 18th, 2020 from the USGS Earth Explorer. https://earthexplorer.usgs.gov/.

After training my model, I applied it to these two images to examine if the slums expanded or gradually disappeared, or stayed the same between the time periods.



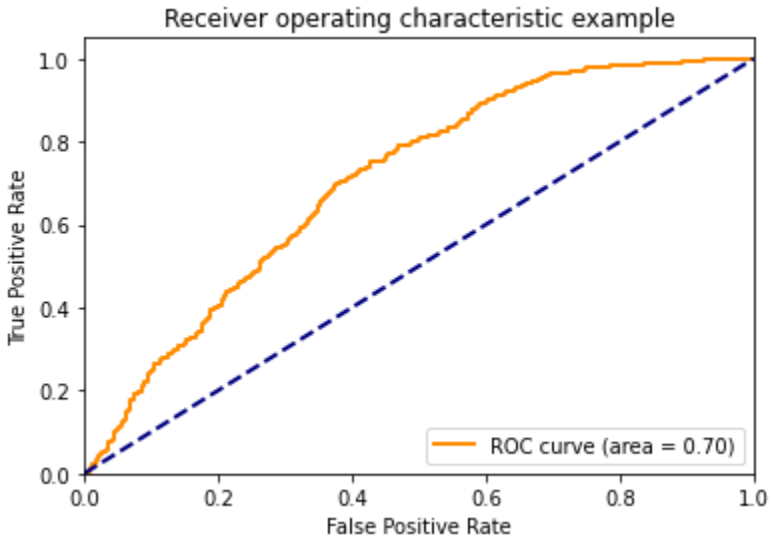
Methods

Support Vector Machine

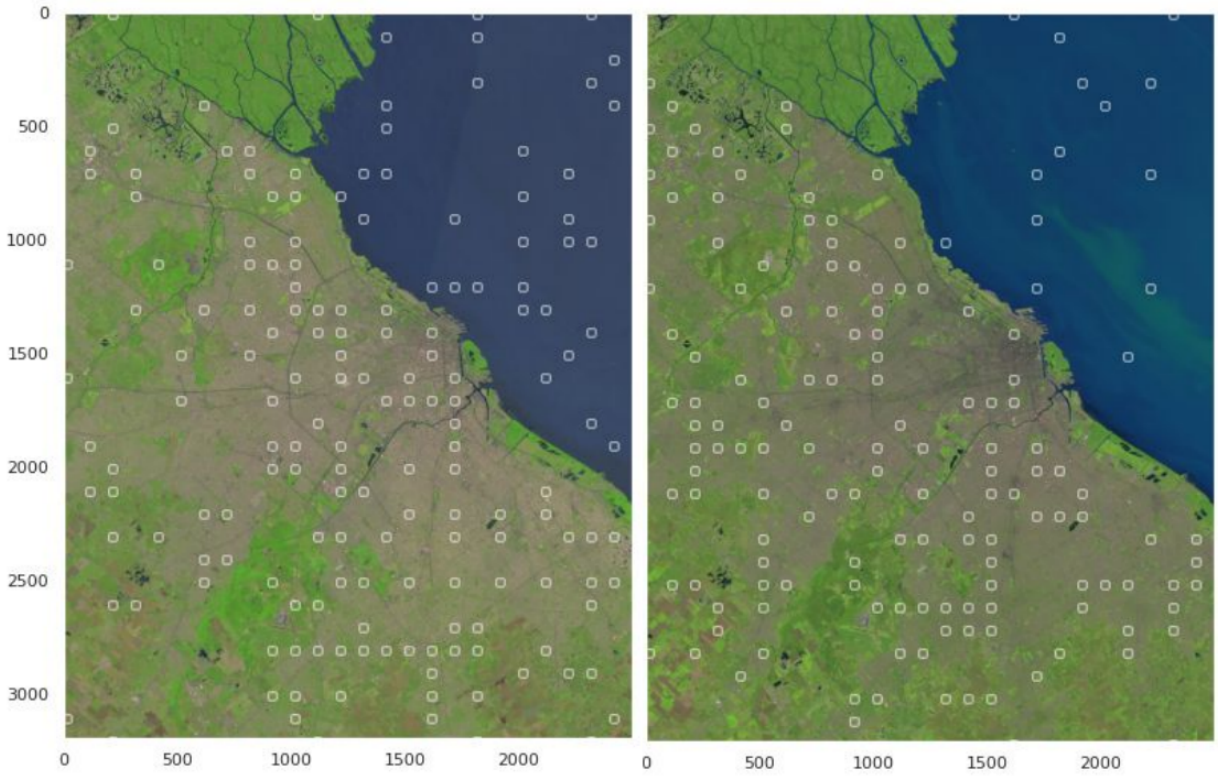
Since this is a classification problem, I chose to start with the simple SVM model. I balanced the number of images in two classes. I then scaled the data to values between 0-255 and combined 3 RGB channels into one gray image array. The data was then split into 80% training and 20% testing.

Using GridSearchCV from the following combinations, kernel: ‘rbf', 'poly', 'sigmoid', gamma: 1e-3, 'scale' and C: 1, 2, 3, 10, 100. The best parameters were ‘rbf’, ‘scale’, and C=1.

These parameters were used in the model to train the whole dataset of the 32x32 pixel images. This model is then applied to the larger satellite images to predict slum locations.



The model had an AUC value of 0.70



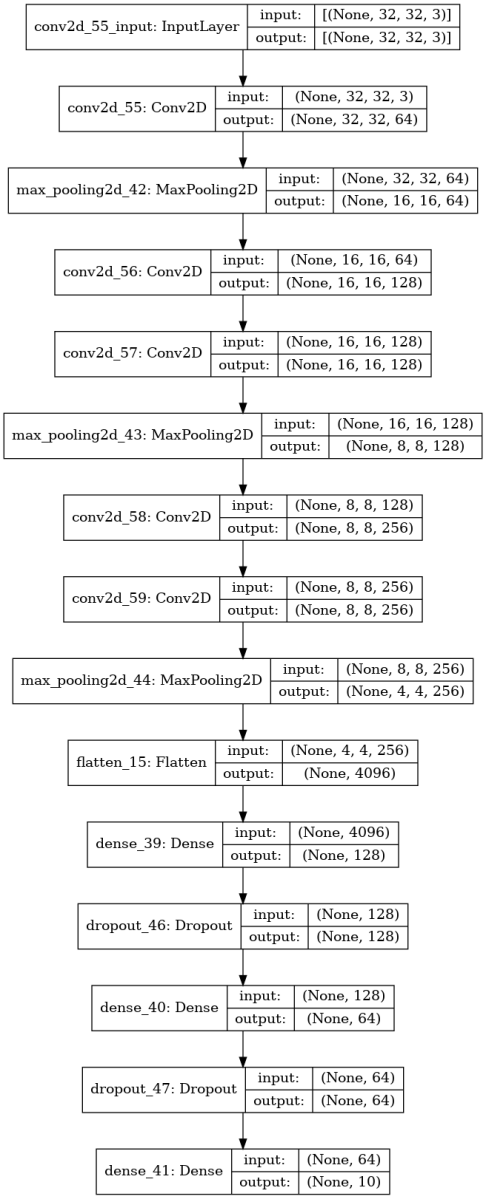
The images above show areas predicted to be slums by the SVM model in and around Buenos Aires. From 2017 on the left and 2020 on the right, the predictions indicated a decrease in the number of slums in the region.

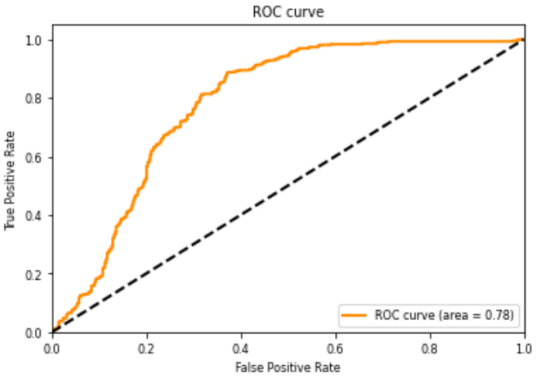
SVM fails to separate ocean images with the land images.

Convolutional Neural Networks

My second model used CNNs to train data which I had split into a training dataset of 60% and testing dataset of 40%. To normalize the data, I divided the image data by the maximum pixel value of the training data. The label data was binarized into two columns, non-slums and slums.

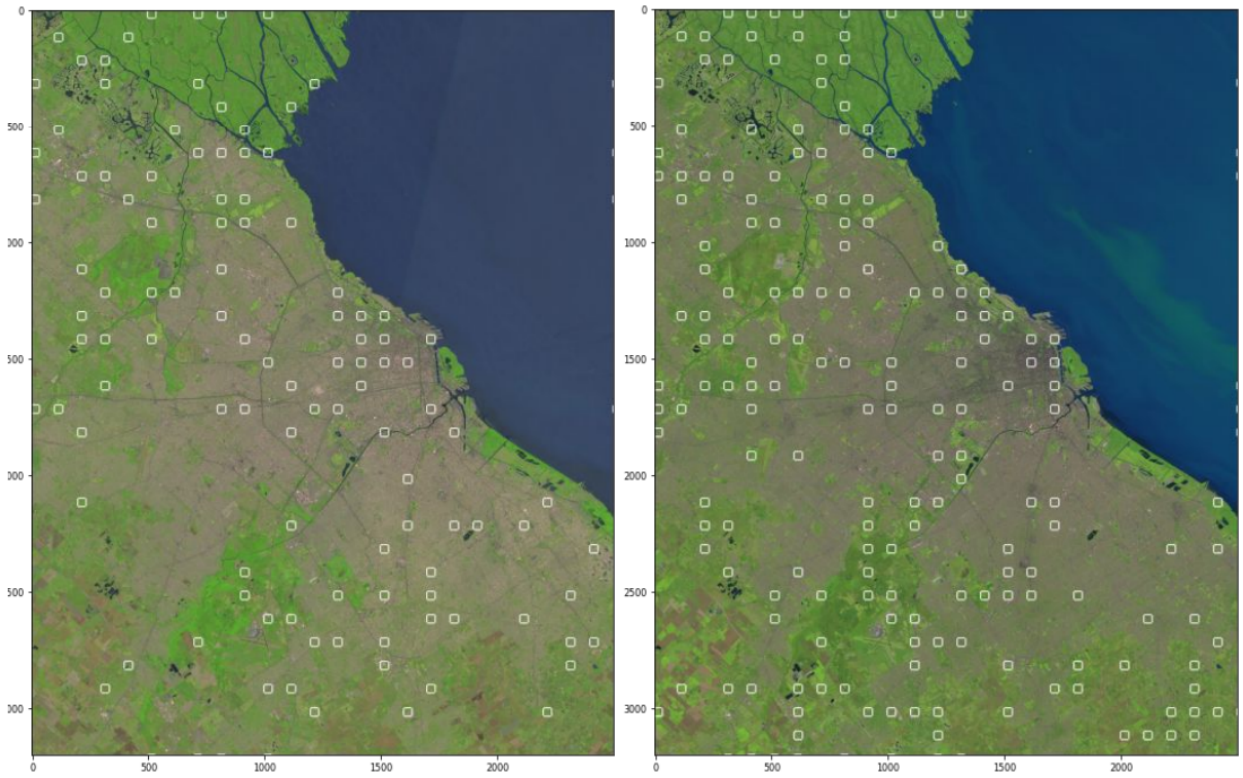
The model uses the optimizer that implements the Adam algorithm and trained the model with 32 as the batch size and 30 as the number of epochs.





The CNN model achieved an accuracy of 0.754, while the AUC of the model is 0.78 as seen above.

After training my model, I applied it to the two satellite images of Buenos Aires and then showed the predicted slums on the larger satellite images.



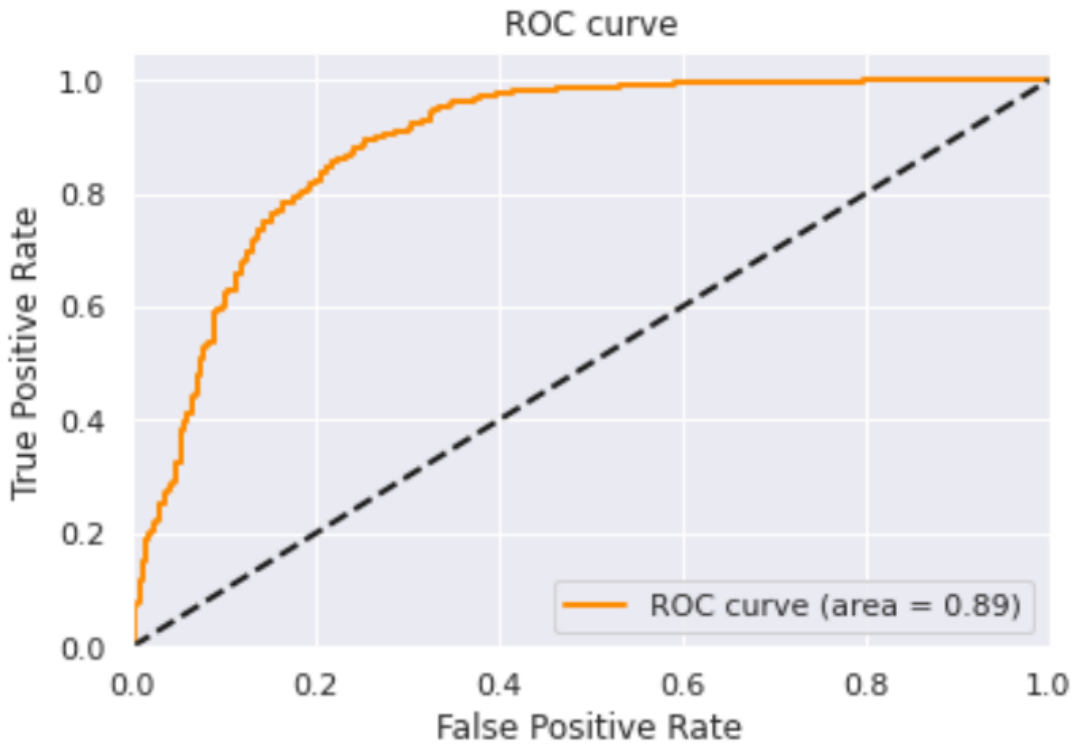
This model successfully identifies the slums in eastern and central Buenos Aires. From 2017 on the left to 2020 on the right, the model indicates an increase in the number of slums in the region.

Residual Networks (ResNet)

In this model, I split the balanced data into 60% training set and 40% testing set. I then normalized the data by scaling the images between a range of 0 - 255, and then divided all pixel values by the maximum values. The label data was binarized into two columns, non-slums and slums.

The model was based on ResNet14v1 to predict whether an image is a slum or not. The model used the optimizer that implements the Adam algorithm and trained the model with batch size of 32 and 80 epochs.

The results of my model were then applied to the downloaded USGS satellite images of Buenos Aires.



Despite the good performance of this model, when I applied it to the big satellite images to find locations of the slums, I could not find any positively labeled prediction which means the model could not find any informal settlements.

Conclusion

The SVM model was able to conduct classification slums and non-slums fairly, although it failed to identify the ocean and predicted many false positives. A better model or diverse data should be included to help the model perform well.

The CNN model did a good job in identifying images that are slums, however in the large satellite images, the model seemed to have predicted more slums in Buenos Aires.

Between CNN and SVM, we see a relatively large difference between the two. SVM predicted a decrease in slums while CNN predicted an increase in slums. The data used is not georeferenced, and we do not have enough training data.

I can conclude that the prediction made by CNN is more reliable and the slums may have increased over the past years.

Next time, I would use high resolution data which will have more useful information (pixels), and find more datasets from sources such as OpenStreetMap to locate the slums.