Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 August 2021 doi:10.20944/preprints202108.0209.v1

Article

Deep learning with uncertainty quantification for slum mapping using satellite imagery

Thomas Fisher †, Harry Gibson †, Gholamreza Salimi-Khorshidi †, Abdelaali Hassaine †, Yutong Cai †, Kazem Rahimi † and Mohammad Mamouei †*

[1] Deep Medicine, Oxford Martin School, University of Oxford, Oxford, United Kingdom.

* Correspondence: mohammad.mamouei@wrh.ox.ac.uk

Abstract: Over a billion people live in slums, with poor sanitation, education, property rights and working conditions having direct impact on current residents and future generations. A key problem in relation to slums is slum mapping. Without delineations of where all slum settlements are, informed decisions cannot be made by policy makers in order to benefit the most in need. Satellite images have been used in combination with machine learning models to try and fill the gap in data availability of slum locations. Deep learning has been used on RGB images with some success but since labeled satellite images of slums are relatively low quality and the physical/visual manifestation of slums significantly varies within and across countries, it is important to quantify the uncertainty of predictions for reliable application in downstream tasks. Our solution is to train Monte Carlo dropout U-Net models on multispectral 13-band Sentinel-2 images from which we can calculate pixelwise epistemic (model) and aleatoric (data) uncertainty in our predictions. We trained our model on labelled images of Mumbai and verified our epistemic and aleatoric uncertainty quantification approach using altered models trained on modified datasets. We also used SHAP values to investigate the how the different features contribute towards the model's predictions and this showed that certain short-wave infrared and red-edge image bands are powerful features for determining the locations of slums within images. Having created our model with uncertainty quantification, in the future it can be applied to downstream tasks and decision-makers will know where predictions have been made with low uncertainty, giving them greater confidence in its deployment.

Keywords: slums; informal settlements; deep learning; machine learning; uncertainty quantification

1. Introduction

Globally, nearly one billion people lives in slums and the figure is estimated to double by 2030. Implicit and explicit social and economic constraints on slum residents result in poor quality of life [1]. In Mumbai, one of the two focus cities for our study, over half of the population live in slums, despite only occupying a small proportion of the urban space. Our other focus city, Bogota, has one of the world's largest slum areas and half of the city's population live in poverty, with over 15% being classed as "very poor". Many of these slum settlements and their populations, however, are neither officially recognised nor mapped accurately by local governments across low and middle income countries (LMICs) due to a substantial lack of technical capacities and resources. Consequently, many slums in LMICs remain under-represented in censuses and surveys and are largely invisible to policy makers, inhibiting targeted provision of infrastructure and welfare for those in greatest need [2].

On top of this data on slums becomes outdated quickly as migration causes informal settlement populations to quickly change [2]. Satellite image data - unlike survey or census data - is readily available in high quality at regular frequent intervals. With

© 2021 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 August 2021 doi:10.20944/preprints202108.0209.v1

machine learning models we can determine slum settlement locations within satellite images to counter the lack of widespread high quality data in this area.

Slum mapping with remotely sensed satellite imagery has been studied extensively and the review paper by [3] has highlighted some important research progress in the area. Compared to more traditional object-based image analysis approaches, machine learning methods have relatively better performance in mapping slums. However, as recommended in the review by [3], purely single pixel-based approaches should be avoided as they ignore surrounding pixel context, indicating that convolutional deep learning models like the U-Net segmentation model [4] could be an appropriate alternative.

The U-Net architecture has seen widespread success in applications to biomedical imagining, geographical mapping, camera feeds from autonomous vehicles and video segmentation [4–7].

Whilst so far convolutional models have been used in this area, the U-Net specifically has not seen application to the slum mapping problem we focus on in this paper, shown in Figure 1. To date, convolutional deep learning models have only been applied to RGB very high resolution (VHR) imagery and only frequentist (i.e. non-Bayesian) approaches have been taken [3,8]. This means that one particular set of model weights is assumed to be optimal rather than taking a Bayesian approach where a posterior distribution of possible weights is considered.

[9] performed satellite image segmentation using convolutional neural networks (CNNs) with a FCN-VGG19 architecture. They also experimented with transfer learning by using model weights from pretraining on the ImageNet database. They found that this transfer successfully increased the Intersection over Union (IoU) score for the model, providing significantly higher scores than when transferring using a model pretrained on images from a different slum from the same country. Some different transfer learned FCN-VGG19 models are investigated further by [10]. Their transfer learning was shown to be more effective than not using transfer learning when predicting on other land cover classes like vegetation. However they found that the best model for predicting a slum class specifically is a FCN-VGG19 using QuickBird VHR data. This model was able to achieve validation recall of 86% and IoU of 77%. We note that calculating either of these metrics require a threshold for the continuous model output in $[0, 1]$ to be chosen resulting in one specific classifier, rather than measuring the classifying power of the model overall in a way that allows for more consistent comparison between models.

The VHR model used by [8] used a DeepLabv3+ architecture. They trained and validated separate instances of the model on different slums that they investigated. They found that their VHR models was generally able to achieve significantly better accuracy and Mean IoU scores than their Canonical Correlation Forest counterparts trained on multispectral data from the corresponding area. It should be noted however that only in two of the eight slum regions investigated did their VHR model achieve accuracy scores of over 90%. They revealed the that the models do not always show good pixel accuracy or mean intersection over union scores for their own validation data. For example, their model trained on Northern Nairobi data only achieves a pixel accuracy of 70% when evaluated on validation data from Northern Nairobi. This level of accuracy on what is a very imbalanced classification problem indicated a relatively weak model.

[11] used a Mask R-CNN segmentation model that utilises transfer learning by pretraining on the COCO image dataset. They used the model to predict the locations of slums in two different images of the same geographical area from the same source taken at two different points in time (2005 and 2018). Both a VHR and a High Resolution (HR) model were investigated and it was found that the VHR model performed better with a strong IoU score of 0.86.

Authors have used different metrics to measure the quality of their models, but all require a single threshold value to be chosen and this poses a significant disadvantage as it does not allow for overall model strength quantification at all threshold values like

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 August 2021 doi:10.20944/preprints202108.0209.v1

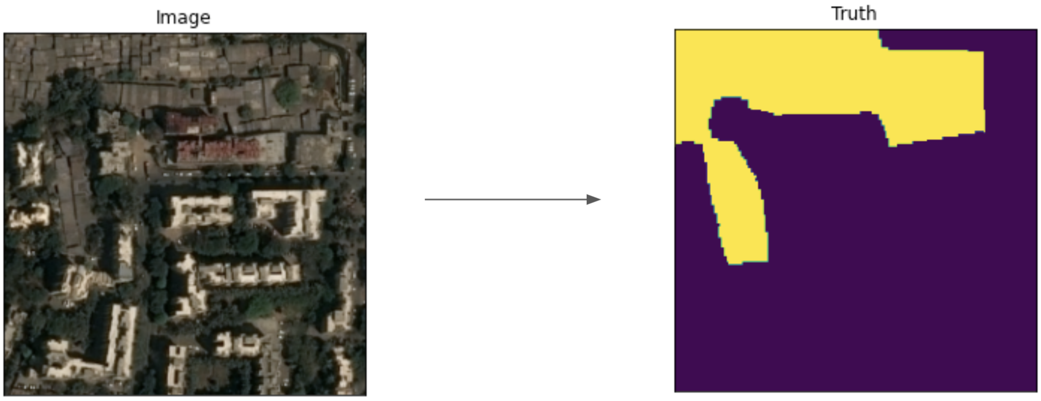


Figure 1. The task we want to train a model to perform well at. It must take in satellite images (left) and a output binary map (right) as to whether or not each pixel is slum (shown in yellow) or not slum (shown in dark purple).

with AUROC or AUPRC. The fact that different metrics have been used does not allow for comparison between models to establish the state of the art model for this task.

It is generally found across all authors that the generalisation of models trained in one slum, when evaluated on a different unseen slum, results in poor performance. This is likely due to a variety of appearances caused by differing building materials, space constraints and geographical features. These cause good model parameters to vary between geographical areas, resulting in poor transferability [12].

This lack of generalisation ability does not allow for reliable applications to downstream tasks that use the model in a broader application. The rarity of quality training data and the interpretability and transferability of models represent pressing problems that require further work in this area. It has been repeatedly emphasised that understanding the levels of uncertainty is an important but as of yet unexplored area of study and represents an important work to be done in order to improve slum model interpretability and deployability [8,12]. If, for example, we were to use the mapping to estimate the number of people living in slums within a geographical region, even a small amount of uncertainty could add up to to become a significant error when calculations are done over the scale of entire cities like Mumbai and Karachi where millions of slum dwellers reside.

Previously, [13] considered spatial (extensional) uncertainties in slum boundaries by asking expert urban scientists with remote sensing knowledge to delineate on maps where they thought the boundaries of slums are. Substantial variations in the predictions made were found among these human experts particularly at the boundaries of settlements, highlighting the importance in objectively quantifying the uncertainty of models for this task. Their study of uncertainty only considers the inter-annotator uncertainty in the data labelling process, whereas we will focus on predictive uncertainties. To the best of our knowledge in the literature there are no interpretable deep learning models for slum mapping which have uncertainty quantification.

In this paper, we aimed to teach an algorithm to be able to produce a map of slums in a satellite image. An example of the model's input and desired output is shown in Figure 1.

Our approach uses a Monte Carlo dropout U-Net model from which we can calculate pixelwise epistemic (model) and aleatoric (data) uncertainties. We trained our model on labelled Sentinel-2 multispectral 13-band satellite images of Mumbai and Bogota and showed that it has good unseen test set performance using AUROC and AUPRC metrics. We verified our epistemic and aleatoric uncertainty quantification approach using altered models trained on modified datasets.

This is the first investigation into either multispectral deep learning models or uncertainty quantification for the slum mapping problem.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 August 2021 doi:10.20944/preprints202108.0209.v1

It is hoped that improving slum mapping can improve our ability to work towards Sustainable Development Goal 11 of inclusive, safe, resilient, and sustainable cities and human settlements [14].

2. Materials and Methods

2.1. Problem statement

We work with a dataset of $N = 2646$ images - target pairs $\mathcal{D} = \{(x_i, y_i) : i = 1, ..., N\}$ from Mumbai and Bogota where the target y_i is a binary image corresponding to and of the same dimensions as x_i. The targets are obtained as ground truth data from local authorities and from previous mapping [8,15].

Figure 1 shows an example image x on the left and the corresponding target y on the right.

We teach a model f to be good at producing targets from given input images so that $f(x_i) \approx y_i$. The model is trained with a pixelwise binary crossentropy loss to make this function f as good at producing targets as possible. We measure the quality of the model predictions using AUROC and AUPRC.

2.2. Model choice: Dropout U-Net

We use the U-Net architecture as initially proposed by [4]. This model has been shown to beeffective at image segmentation tasks in a wide variety of applications.

The U-Net consists of an encoder-decoder architecture. The encoder part downsamples the image using convolutions and max-pooling operations to learn a compressed representation of the image in feature space that contains the essence of its contents. The decoder part takes this compressed image and unpacks the features using upsampling and transposed convolutions to output a segmentation of the same size as the input image. This means that every pixel is labelled as slum or not-slum in our case. However rather than classifying each pixel based purely on its feature values alone, the convolutional part of the model uses the context of surrounding pixel feature values also during the convolution process. Using an approach like this which is not based on individual separate pixel classification is important as emphasised in [3].

We used a normaliser trained on the training data to help the gradient descent process converge more quickly during training.

We use Monte Carlo dropout [16] between every layer of the architecture which intuitively randomly kills off a small number of weights in the model. Each of these different randomly dropped-out models make a slightly different prediction, which simulates the process of sampling from the distribution on the weights. Dropout also prevents overfitting during training [17]. Monte Carlo dropout is explained further in Supplementary Material and [17].

We use 500 dropout models in our approach to ensure that we sample a relatively wide range of models from the distribution. The dropout rate for all the of models was 0.25.

We used the popular Adam optimiser [18] with initial learning rate 0.001. The Adam optimiser is adaptive and so we do not expect results to change significantly when altering this value. We trained our model for 100 epochs with a batch size of 128.

We use the common binary crossentropy loss as outlined in Supplementary Material.

2.3. Training and Data Configuration

We use two locations with slum maps in this work. The first is Mumbai (India), where in 2011 PK Das & Associates produced an award-winning map of the city's slums. The second is Bogota (Colombia), where in 2018 Gram-Hansen produced a slum map of the city and used it to train their own models [8].

We used 10 meter Resolution 13-band multispectral imagery from the Sentinel-2 satellite through the Descartes Lab platform. All of the imagery data is freely available

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 August 2021 doi:10.20944/preprints202108.0209.v1

through the European Space Agency. This satellite was launched in 2015 and we use annual images between 2015 and 2020. We tiled the Mumbai and Bogota regions, using square tiles width 64 pixels with 2 pixels of padding. Details of the dataset creation process and the areas of interest can be found in the Supplementary Materials.

In this paper we combine the Sentinel-2 data from both Bogota (1311 image tiles) and Mumbai (1335 image tiles) to produce a large multi city dataset containing 2646 image tiles. More information about data preprocessing can be found in the Supplementary Material.

We used 10% of the dataset for validation and 10% of the dataset as a separate test set. The other 80% was used for training. The data was randomly shuffled before the data split was performed.

2.4. Calculation of Uncertainty

[19] provides ways of approximating the two different types of uncertainty for binary classification models using predictive entropy and mutual information.

A more intuitive approach is obtained by following [20]. In our binary classification case where the dropout model with index t assigns probability $\hat{p}_t$ to the positive class as its final output for a pixel in a particular input image, their approximations simplify to calculating :

$$\text{Aleatoric Uncertainty} \approx \frac{1}{T}\sum_{t=1}^{T}\hat{p}_t(1-\hat{p}_t) \tag{1}$$

$$\text{Epistemic Uncertainty} \approx \frac{1}{T}\sum_{t=1}^{T}(\hat{p}_t-\bar{\hat{p}})^2 \tag{2}$$

where $\bar{\hat{p}} := \frac{1}{T}\sum_{t=1}^{T}\hat{p}_t$ at every pixel.

The epistemic uncertainty here quantifies the uncertainty as called for in [12].

For more explanation about the two types of uncertainty used in this paper, see the Supplementary Material.

2.5. Metrics

As the dataset is highly imbalanced with many more not-slum pixels than slum pixels, accuracy alone is not a good measure of the performance of a model. Only about 10% of the pixels in the dataset are slum so a naive classifier that always assigns pixels to the “not slum” category has an accuracy 0.9.

Some authors use use intersection over union or individual precision and recall scores for a specific output threshold output to measure the quality of their algorithm. However these scores do not indicate how powerful overall the model is across all threshold values.

We use two Area Under Curve (AUC) metrics to measure the performance of our model, which are the standard way of comparing binary classifiers. The first is Area Under the Receiver Operator Characteristic (AUROC). This calculates the area under the curve of True Positive Rate against False Positive Rate at all possible threshold settings. The second is Area Under the Precision-Recall Curve (AUPRC). This calculates the area under the curve of Precision against Recall at all possible threshold settings. This metric is typically more informative than AUROC for imbalanced datasets and does not give a false sense of high performance on imbalanced classification tasks [21].

Both of these two metrics measure the overall predictive power of models without a threshold needing to be chosen for classification. The AUPRC can be interpreted as the average Precision obtained for all threshold values. For both metrics, values closer to 1 indicate a stronger model. The

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 August 2021 doi:10.20944/preprints202108.0209.v1

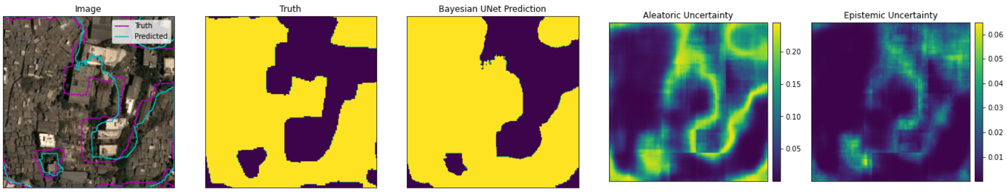


Figure 2. Predictions and uncertainty for different tiles when using the U-Net model. Each row shows the tile, the ground truth label without masking (yellow denotes slum and dark purple denotes not slum), the Bayesian U-Net prediction along with the aleatoric and epistemic uncertainties in these predictions. Note that we have shown higher resolution RGB images in the first column whilst lower resolution multispectral images were actually used by the model.

3. Results

We used 10 meter multi-spectral satellite images from the European Space Agency Sentinel-2 satellite along with pre-existing high quality slum maps from [8] and [15] from which we produced pixel-wise labels of where slums are located within the satellite images. These maps appear visually correct at when inspected in close detail. Details about the dataset creation process can be found in the Supplementary Material.

We split 64x64 pixel images randomly to train (80%), validation(10%) and a seperate unseen test set(10%). The dropout U-Net model was trained using binary crossentropy loss to predict the the correct segmentation of the image into pixels that corresponds to a slum and pixels that correspond to non-slum as shown in Figure 1. The quality of the model predictions was measured as based on the pixel-wise test AUROC and AUPRC, which measure the classifying power of the model overall in a way that allows for consistent comparison between models.

3.1. Training History and Test Set Performance

After 100 epochs of training our multispectral dropout U-Net model achieves good validation AUROC and AUPRC scores of around 0.98 and 0.94 respectively. Plots of the training history can be found in Supplementary Material. Testing on our separate held out test set gives AUROC and AUPRC scores of 0.90 and 0.84 respectively. This indicates a strong model is obtained through this training process. A model with performance as good as this is appropriate for use on complex downstream tasks.

3.2. Visualisation of Predictions

Figure 2 shows a prediction made by the Bayesian model along with the aleatoric and epistemic uncertainty at each pixel for an example input images. An optimal threshold of 0.59 was found to maximise the F1 score and was used to determine whether model outputs in $[0, 1]$ were classified as "slum" (above the threshold) or "not-slum" (below the threshold).

Note that we have shown a Airbus Pleadies RGB 1 meter resolution image (i.e. not the multispectral Sentinel-2 10 meter resolution image) in the first column of each figure. This is because multispectral and not possible for humans to visualise as they contain 13 bands of colour and we can only see 3. Also the RGB components of these multispectral image are at a lower resolution making the urban cover types difficult to see.

We found that generally aleatoric uncertainty is typically very high (in fact, sometimes getting close to its maximum value of 0.25) at the boundary of the segmentation, but is generally low elsewhere. It is reassuring that we observed that there was generally not high aleatoric uncertainty towards the center of regions labelled as slum. This indicates that the boundary distinctions are the more challenging places to classify, perhaps with larger ranges of possible behaviour on the edge of the settlements with either greenery or roads or other types of buildings neighbouring the slums. More examples of model predictions and uncertainty quantification can be found in the Supplementary

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 August 2021 doi:10.20944/preprints202108.0209.v1

Material. As we show in the histogram plots in the next section, the epistemic uncertainty does not often reach particularly high values.

Overall these examples show that our model predictions align with what we would expect given the validation metric values whist also showing the difficulties with uncertainty emphasised by [13].

3.3. Uncertainty distribution

We demonstrate the validity of our uncertainty quantification approach by considering two examples which simulate using inaccurate and smaller training datasets showing us how the uncertainties change with lower quality and lower quantity training data scenarios respectively. We used three variants of our model. The first "Standard" model is our dropout U-Net as described in the Methods section. The "Class Flip" model was trained on the same training data but with 10% of the "slum" pixels flipped to the "not-slum" category at random, simulating what happens when slum regions are missed in the labelling process. The "Half Data" model was trained using only half as much training data as the Standard model, simulating what happens when training data is scarce. The intuition behind investigating these two altered models alongside the standard model is that we expect to see higher epistemic uncertainty for the Half Data model and higher aleatoric uncertainty for the Class Flip model both when compared with the Standard model. These altered models therefore allow us to validate our uncertainty quantification approach.

In Figures 3 and 4 we plot the uncertainty of the model in each prediction for pixels in the test set. This is plotted against a logarithmic frequency of occurrence in base 10 on the vertical axis. We note going forward that the approximations obtained for both types of uncertainty derived in Section 2.4 are capped at 0.25 for each pixel as this is a binary segmentation problem.

Using our uncertainty quantification approach we can determine how much uncertainty is introduced in these different scenarios that practitioners might find themselves in if different data was available to them. Whilst we found that AUPRC was not noticeably decreased for these other two inferior (by construction) models, the uncertainty quantification can show that they are less good models.

Figure 3a shows that there is no substantial difference between using the Standard and Class Flip models on the epistemic uncertainty of slum pixel classifications. However using the Half Data model increases the epistemic uncertainty significantly; the Half Data histogram has much more mass distributed at higher epistemic uncertainty values above 0.1 than the other two models.

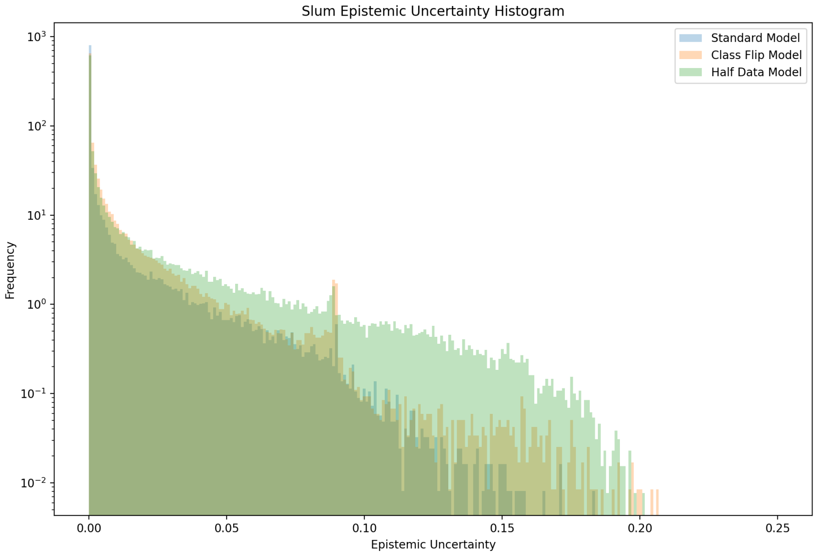
Figure 3b shows us that all three models have essentially the same epistemic uncertainty on the not-slum pixels in the test set. This is as the proportion of pixels in the not-slum category is much higher than the slum category and so the Half Data model is still exposed to sufficient data in the not-slum category. The Class Flip model is trained with some pixels incorrectly relabelled from slum to not-slum, hence this Class Flip model only has slightly increased not-slum epistemic uncertainty in the 0 to 0.05 range compared to the other two models.

Overall the epistemic uncertainty is substantially higher when using the Half Data model. Using our dropout model this change is measurable at the pixelwise level by plotting these histograms. Figure 4 shows that there is little change in aleatoric uncertainty between the Standard and Half Data models.

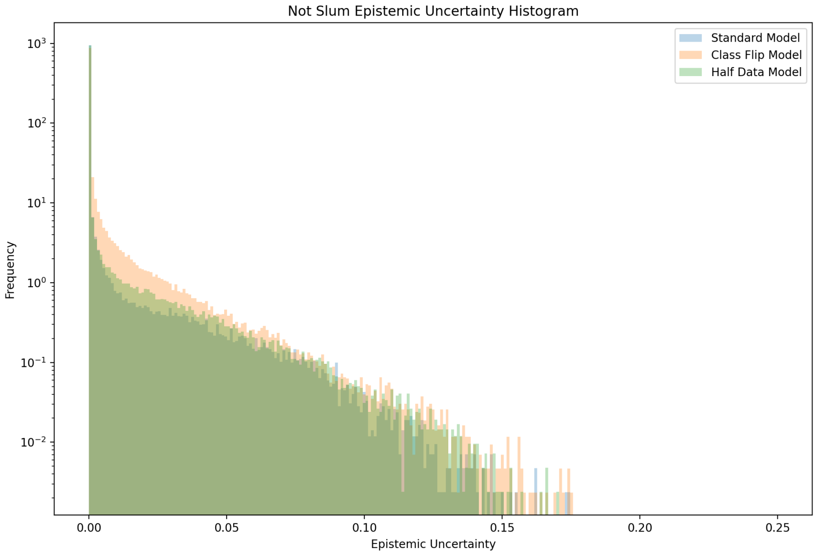
The increased epistemic uncertainty and no change in aleatoric uncertainty of the Half Data model compared to the Standard model is in agreement with the general rule that reduced training set size will result in lower certainty in model parameter values hence increasing the epistemic uncertainty.

Figure 4a shows that whilst the aleatoric uncertainty histograms of the three models on slum pixels in the test set are quite similar, the Class Flip model has the highest mass of the three models at higher uncertainty values.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 August 2021 doi:10.20944/preprints202108.0209.v1



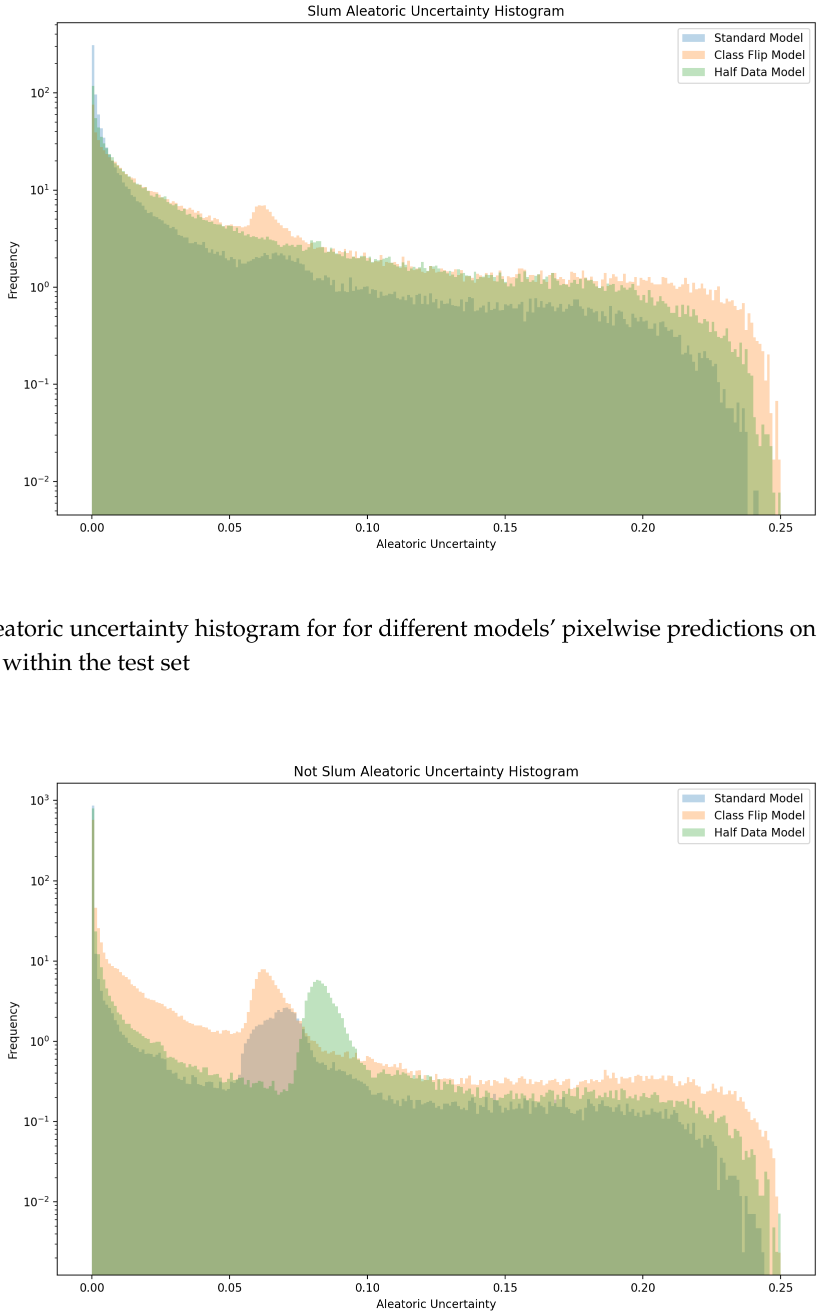
(a) Epistemic uncertainty histogram for different models' pixelwise predictions on the slum pixels within the test set



(b) Epistemic uncertainty histogram for different models' pixelwise predictions on the not-slum pixels within the test set

Figure 3. Density histograms with logarithmic scale for epistemic uncertainty on slum and not slum test set pixels for the three different models

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 August 2021 doi:10.20944/preprints202108.0209.v1



(a) Aleatoric uncertainty histogram for for different models' pixelwise predictions on the slum pixels within the test set

(b) Aleatoric uncertainty histogram for for different models' pixelwise predictions on the not-slum pixels within the test set

Figure 4. Density histograms with logarithmic scale for aleatoric uncertainty on slum and not-slum test set pixels for the three different models

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 August 2021 doi:10.20944/preprints202108.0209.v1

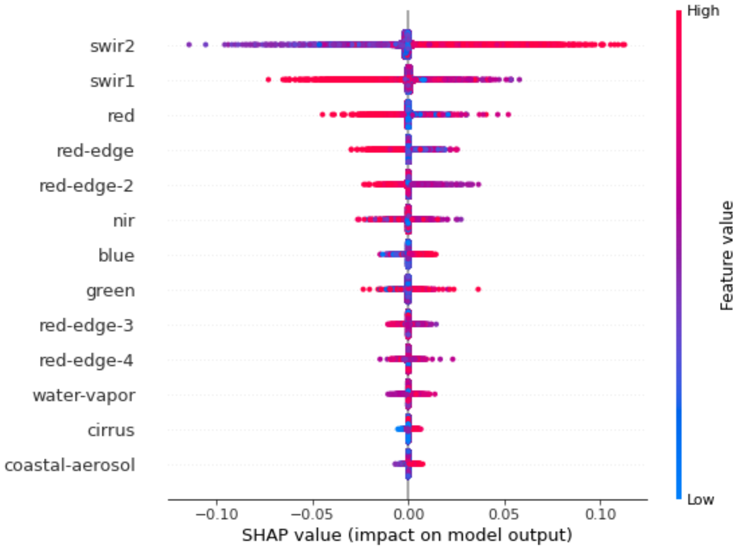


Figure 5. SHAP value summary plot for pixels in the 2015 Sentinel-2 Mumbai test dataset from our U-Net model. Each dot represents a pixel. Features are ranked vertically in descending order of feature importance. The horizontal position of the dots represent the feature impact on the prediction using the SHAP value. The colour of the dot represent the whether the feature value is high or low for that pixel.

Similarly, Figure 4b shows that whilst the aleatoric uncertainty histrograms of the three models on not-slum pixels in the test set are similar, again the the Class Flip model has the highest mass of the three at higher uncertainty values.

We conclude that the Class Flip model has the highest aleatoric uncertainty of the three models. Figure 3 shows that there is little change in epistemic uncertainty between the Standard and Class Flip models. Our observed increase in aleatoric uncertainty and no change in epistemic uncertainty the Class Flip model when compared to the Standard model agrees with the general rule that poorer quality of data results in increasing the this type of uncertainty.

The plots the distribution of epistemic and aleatoric uncertainty in Figures 3 and 4 show that both uncertainties have substantial positive skew. These plots reassure us that the vast majority of pixels have both epistemic and aleatoric uncertainty below 0.05 (recalling that the histograms are plotted with a log scale). The epistemic uncertainty histogram shows that there is negligible density between 0.1 and 0.25, informing us that significant uncertainty in the model weights is not a problem. The aleatoric uncertainty histogram shows that there still some density between 0.1 and 0.25, representing the inherent irreducible uncertainty in these predictions when this dataset is being used.

We can see this uncertainty distribution plot as one of the main strengths of the Bayesian approach: it is reassuring that the model uncertainty is high on only a very small number of test set pixels, (as shown on the left hand side of Figure 3), quantification of which is not possible in a frequentist approach.

3.4. Model Interpretability

3.4.1. Mumbai

Usually, one of the main disadvantages of using deep learning architectures like U-Net is the lack of model interpretability available. We overcome this by calculating the SHAP values [22] of our model on test set pixels to interpret our U-Net model and gain insight into how our model associates different input feature values with outputs. By using SHAP values we obtain information about both feature importance and the influence of feature values on predictions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 August 2021 doi:10.20944/preprints202108.0209.v1

We can read off the five most important features from Figure 5 as swir2, swir1, red, red-edge and red-edge-2. We can see that there is a trend shown in Figure 5 that higher values of swir2 intensity generally have positive SHAP values indicating an association between the model being more likely to predict “slum” and high values in the swir2 band. This is because the model outputs a sigmoid value in $[0, 1]$ where values closer to 0 indicate higher probability of “not-slum” and values closer to 1 indicate higher probability of the “slum” class.

Higher values of swir1 intensity generally have negative SHAP values indicating an association between the model being more likely to predict “not-slum” and high values in the swir1 band.

We see that the other three of the five most important features (red, red-edge and red-edge-2) have similar impact on the model output to swir1: higher values in these bands are associated with lower model output which represents the pixel being more likely to be in the “not-slum” class.

Our findings were robust to different subsets of the test-set being used for the SHAP value analysis.

Figure 5 agrees with Kotthaus et al [23] that information about the material composition of settlements can be determined with data resolved to the wavelength level, with reflectance in the visible to short-wave infrared region being powerful predictors remote sensing-based land use classification. In our setting this can be interpreted as slums having a distinctive signature in these image bands different to formal housing, which is picked up by the model.

We note that whilst they are not the least important features, the green and blue colour bands do not have large feature importance in our LR model. This indicates that VHR RGB imagery alone is not nearly as powerful as multispectral imagery at this resolution for this problem. Our findings that swir bands being most powerful features align with the [24] use of these features in their Normalized Difference Built-Up Index (NDBI) for distinguishing built-up areas.

3.5. Generalisation to other slums

We have so far performed all of our analysis using models trained and validated on images of Mumbai and Bogota. When testing on images of other slums we found that the performance is poor. Other authors have noticed the same; the transfer of a model trained on one slum to testing on a different slum results in poor performance [12]. We present a mathematical argument for why this is the case is outlined using the Maximum Mean Discrepancy in the Supplementary Material.

4. Discussion

This paper is a proof of concept for interpretable deep learning models for slum mapping from satellite images with uncertainty quantification. We emphasise that this approach is novel in this area of application and our model is quite different to what has been used previously in the literature.

The unseen test set performance obtained in this paper indicates that there is significant scope for use of this model in downstream applications in Mumbai and Bogota. An example of such an applications might be slum monitoring where changes in slum geography are tracked over time to help inform policy decisions.

We experimented with different model variants to measure the impact separately of reduced quality and reduced quantity of training data. We observed the changes in aleatoric and epistemic uncertainty distributions in predictions and confirmed that the uncertainty quantification works as it should.

We showed using SHAP values that the model’s predictions are interpretable and discovered that the model uses certain multispectral image bands (particularly short wave and near infrared bands) as the most powerful features for segmentation, picking up on a slum signature in these features.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 August 2021 doi:10.20944/preprints202108.0209.v1

Our work has provided unique, novel contributions in a number of ways. We have indicated the feasibility and importance of using multispectral images for slum mapping. Previously these types of images had not been used in conjunction with deep learning models. We have, for the first time created a slum mapping model which quantifies aleatoric and epistemic uncertainty in its predictions. Previously all mapping models were purely frequentist models where uncertainty quantification for predictions was not possible. With our uncertainty quantification users know when the model is more confident (low epistemic uncertainty) and where the model is has seen similar data before (low aleatoric uncertainty). This gives users of the model reassurance in the information the model outputs when decisions are being made regarding infrastructure or policy changes in slum areas. Without using uncertainty quantification the user might not be able to detect poorer predictive certainty as the AUROC and AUPRC scores of the variant models we investigated were not dissimilar from the Standard model.

Clearly, our approach has some limitations.

Firstly, we are only using satellite images to determine slum from not-slum. There may be differences between residences which are indistinguishable from the remote sensing imagery, causing one area to be slum and the other to be not-slum according to UN definitions.

Secondly, we only have access to slum maps from one point in time (2011 for Mumbai and 2018 for Bogota) and do not have satellite images of Mumbai from 2011. We used composite images with images from different months of the year combined together into single images for each year in order to improve image quality. This requires us to assume that there is not significant change in the slum areas between these different months of the year.

We think that future satellite imagery products should ensure consistency between imaging equipment and angles which would facilitate easier transfer of machine learning between image collections. Our model only works on images from the satellite sensors that it has been trained on. We would like to see work towards a more universal model that takes as input images of different origins, resolutions and images bands.

Finally, no target truth slum map is likely to be perfect. Without going into every residence we cannot be absolutely sure that the map is entirely complete. There are slightly differing definitions of what counts as slum and usually only sparse survey data rather than census data exists for these areas [2]. We have assumed that we can overcome this difficulty by training models on large amounts of data to get reasonable performance.

As we have shown that our model is of high quality - it is able to achieve high levels of performance on unseen test data whilst proving the added benefits of interpretability and uncertainty quantification. This means that our models is appropriate for deployment to all kinds of downstream applications such as slum monitoring over time and combining other kinds of data to make predictions on other variables like environmental health. We would like to see future papers using this model in these kinds of tasks in order to provide better data for policy makers to inform their decisions.

We think that it is important for researchers to use the same metrics allowing for model performance comparison to make the top performing start the art model clear. As explained in Section 5.5, the use of AUPRC is much more appropriate for this problem than the metrics used by researchers previously, and a consistent

We also think that the questions of model transferability brought up by [3], [9] and [12] is an important area of study which yet remains unsolved.

Funding: This research is part of the Informal Cities Programme funded by the Oxford Martin School, University of Oxford.

Acknowledgments: In this section you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 August 2021 doi:10.20944/preprints202108.0209.v1

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wekesa, B.W.; Steyn, G.S.; Otieno, F.A. A review of physical and socio-economic characteristics and intervention approaches of informal settlements. *Habitat International* **2011**, *35*, 238–245. doi:10.1016/j.habitatint.2010.09.006.
2. Lucci, P.; Bhatkal, T.; Khan, A.; Berliner, T. What works in improving the living conditions of slum dwellers.
3. Kuffer, M.; Pfeffer, K.; Sliuzas, R. Slums from space-15 years of slum mapping using remote sensing, 2016. doi:10.3390/rs8060455.
4. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.
5. Krestenitis, M.; Orfanidis, G.; Ioannidis, K.; Avgerinakis, K.; Vrochidis, S.; Kompatsiaris, I. Oil spill identification from satellite images using deep neural networks. *Remote Sensing* **2019**, *11*, 1762. doi:10.3390/rs11151762.
6. Tran, L.A.; Le, M.H. Robust u-net-based road lane markings detection for autonomous driving. Proceedings of 2019 International Conference on System Science and Engineering, ICSSE 2019. Institute of Electrical and Electronics Engineers Inc., 2019, pp. 62–66. doi:10.1109/ICSSE.2019.8823532.
7. Liu, H.; Jiang, J. U-Net Based Multi-instance Video Object Segmentation **2019**. [1905.07826].
8. Gram-Hansen, B.; Helber, P.; Varatharajan, I.; Azam, F.; Coca-Castro, A.; Kopackova, V.; Bilinski, P. Mapping Informal Settlements in Developing Countries using Machine Learning and Low Resolution Multi-spectral Data. *Proceedings of AAAI/ACM Conference on AI, Ethics, and Society* **2019**.
9. Stark, T.; Wurm, M.; Taubenböck, H.; Zhu, X.X. Slum Mapping in Imbalanced Remote Sensing Datasets Using Transfer Learned Deep Features. 2019 Joint Urban Remote Sensing Event (JURSE), 2019, pp. 1–4. doi:10.1109/JURSE.2019.8808965.
10. Wurm, M.; Stark, T.; Zhu, X.X.; Weigand, M.; Taubenböck, H. Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing* **2019**, *150*, 59–69. doi:10.1016/j.isprsjprs.2019.02.006.
11. Maiya, S.R.; Babu, S.C. Slum segmentation and change detection: A deep learning approach, 2018, [1811.07896].
12. Gevaert, C.M.; Kohli, D.; Kuffer, M. Challenges of mapping the missing spaces. 2019 Joint Urban Remote Sensing Event (JURSE), 2019, pp. 1–4. doi:10.1109/JURSE.2019.8809004.
13. Kohli, D.; Stein, A.; Sliuzas, R. Uncertainty analysis for image interpretations of urban slums. *Computers, Environment and Urban Systems* **2016**. doi:10.1016/j.compenvurbsys.2016.07.010.
14. Kuffer, M.; Wang, J.; Nagenborg, M.; Pfeffer, K.; Kohli, D.; Sliuzas, R.; Persello, C. The scope of earth-observation to improve the consistency of the SDG slum indicator, 2018. doi:10.3390/ijgi7110428.
15. PKDas. PK Das Slum Map of Mumbai. http://www.pkdas.com/maps/3-Mumbai's-Slums-Map.pdf, 2011.
16. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, 2016, ICML'16, p. 1050–1059.
17. Duerr, O. *Probabilistic Deep Learning: With Python, Keras and TensorFlow Probability*; Manning Publications Company, 2020.
18. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization, 2017, [arXiv:cs.LG/1412.6980].
19. Gal, Y. Uncertainty in Deep Learning. PhD thesis, University of Cambridge, 2016.
20. Kwon, Y.; Won, J.H.; Kim, B.J.; Paik, M.C. Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation. *1st Conference on Medical Imaging with Deep Learning* **2018**. doi:10.1016/j.csda.2019.106816.
21. Davis, J.; Goadrich, M. The relationship between precision-recall and ROC curves. ACM International Conference Proceeding Series; ACM Press: New York, New York, USA, 2006; Vol. 148, pp. 233–240. doi:10.1145/1143844.1143874.
22. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions, 2017.
23. Kotthaus, S.; Smith, T.E.; Wooster, M.J.; Grimmond, C.S. Derivation of an urban materials spectral library through emittance and reflectance spectroscopy. *ISPRS Journal of Photogrammetry and Remote Sensing* **2014**, *94*, 194–212. doi:10.1016/j.isprsjprs.2014.05.005.
24. Zha, Y.; Gao, J.; Ni, S. Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *International Journal of Remote Sensing* **2003**, *24*, 583–594. doi:10.1080/01431160304987.