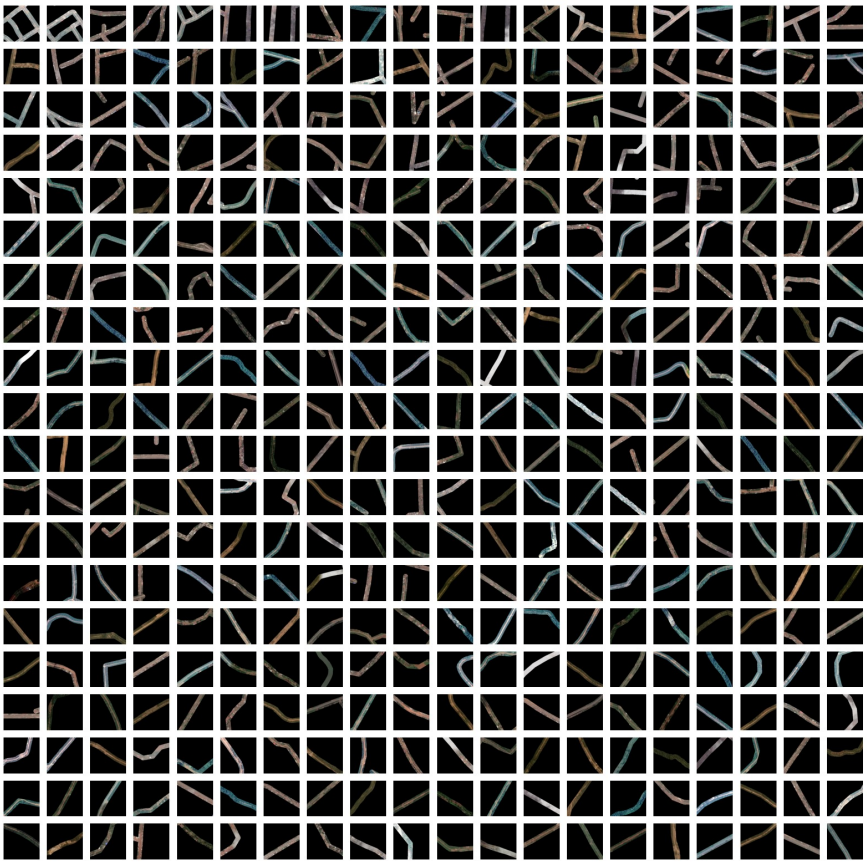
# Leveraging Satellite Imagery to Assess Road Quality in the Democratic Republic of the Congo

Miguel Conner Bonmatí
Ramón Talvi Robledo
Dominik Wielath



12 June 2023

Data Science Methodology - Masters Thesis

Barcelona School of Economics

# Acknowledgements

We would like to thank Hannes Müller, our supervisor for this thesis as well as Mathilde Lebrand, Jevgenijs Steinbuks, Jay Taneja, Aggrey Muhebwa, Peer Schouten and Nuri Bae for the insightful debates about road quality assessment in the DRC.

We further want to acknowledge financial support from the World Bank Group.

# Table of Contents

# Abstract

We attempt to build a road quality classifier to detect bad roads using satellite imagery in the province of Sud-Kivu in the Democratic Republic of the Congo (DRC). We model our approach on existing literature but make a few deviations. Using 60 cm/pixel resolution from Google Earth, paired with 100 m IRI road quality data for Liberia, we train a CNN (EfficientNetV2) that performs with an accuracy of 47% for 5-classes and 80% for 2-classes (AUC: 0.75). Using maps of road types for Liberia and the DRC, we then establish a connection between the model trained in Liberia and road quality in the DRC. We find that our methods seem to work well given the many limitations of the project. We note these limitations and suspect that more standardized higher-quality imagery would be helpful to achieve better results.

# 1. Motivation

Roads are critical pieces of infrastructure that are often taken for granted. In the Democratic Republic of the Congo (DRC), roads have an outsized importance on the safety and security of it's population. In this section, we aim to convey the geographical, political, and sociological factors that motivate the development of an accurate road quality detector for the eastern DRC.

**Geography and Soil Type**

The geographic characteristics of the DRC make road construction and maintenance very difficult. The DRC is a large country; the second largest country in Africa spanning just over 2.3 million square km. Its population of about 112 million people is widely dispersed throughout, often separated by large portions of dense rain forest and criss-crossing rivers. The equatorial placement and humidity of the tropical climate create conditions for significant rainfall. Indeed, the rainy season lasts from September to May and often drops over 2000 mm of rain per year in some areas.

Perhaps the most important feature of the geography in terms of road building is the thick, slippery, clay-like dirt called vertisol, which covers much of the country. Schouten et al. (2022) explain:

> "Vertisols form where sediments wash into the depressions of undulating landscapes, where humans typically chart their course to reduce the friction of terrain. They heavily swell with rain, saturate quickly to give rise to viscous puddles and glutinous gullies, and shrink into hard, cracked earth when drying. In tropical conditions, implying humidity for most of the year, roads become, in a very

literal sense, 'sticky' infrastructures, pulling on wheels, feet and shoes (Duffield, 1970)."

Building roads under these conditions is difficult, but maintaining them is even more so. Traditionally, any long distance movement of migration before the 20th century was limited to the dry season. Since then, many have struggled to maintain a passable road network.


Figure 1.1: Travel can be quite difficult on DRC roads. (Schouten et al., 2022)

### Brief History

Congo gained its independence from Belgium in 1960, but its impressive road network shrunk by 60% almost immediately. The colonizer's solution to dealing with the vertisol was to require all able bodied men to spend a month every year maintaining the roads. Now a free country however, nobody could be required to do this work, and nobody could pay for someone to do it either.

For political reasons, the dictator Mobutu, who ruled from 1965 to 1997, was wary of creating a well-connected road network, though he did get many major donations to maintain and build roads. But by 1985 only 15% of the road network of 1960 was passable.

**Infrastructure Investment**

In 2022 the World Bank approved a plan to spend $500 million to strengthen transport and connectivity in the DRC (World Bank Group - Press Release, 2022).[1] One of the plan's key points focuses on upgrading and paving 440 km of weather-resistant roads in the Eastern part of the country (North Kivu and Kasai provinces in particular). This is not the first major investment of this kind, as Schouten et al. (2022) write: "The World Bank poured billions of dollars into the country's [DRC] transport network in a Sisyphean procession of projects that brought neither durable development nor state authority." It is pivotal to assess road quality before and after projects are carried out, to guarantee the success of infrastructure investments and the most effective use of resources. Monitoring road quality on a large scale on the ground is costly and impossible in conflict-prone areas such as Eastern Congo, even more so if the evaluation needs to be updated regularly. With this project, we aim to evaluate the prospects of leveraging satellite imagery to carry out this large-scale evaluation at a relatively low cost.

[1]The authors of this thesis were funded by the World Bank to carry out this project.

# 2. Literature Review

This section summarizes the literature on road quality assessment leveraging satellite imagery and machine learning. It focuses on the data sources and methods used and the achieved results.

## 2.1 Measures of Road Quality

When discussing road quality assessment, it is crucial first to define road quality and identify an established measure of road quality serving as ground truth to which we can relate the performance of our approach. In this regard, the literature is converging upon a standard, the vertical displacement of a vehicle as it travels along a road. The gold standard of measuring this vertical displacement is the International Roughness Index (IRI).

The IRI is a measurement that was created by the World Bank in 1986 to measure the "bumpiness" of any road, from airplane runways to jungle dirt roads. To obtain IRI measurements a device measuring vertical displacement is calibrated to a specific vehicle. This vehicle then travels along the road at a given speed to continuously measure vertical displacement per kilometer traveled (m/km). Depending on the data sources, this continuous displacement is aggregated and averaged over segments of different lengths. Therefore, IRI measurements exist at different granularity levels, with average IRI values, for example, for every ten or every 100 meters. These values are usually stored in shapefiles, which combine the IRI information with the precise geometric location of the road segment from which they originate. While IRI is the gold standard of road quality evaluation, it is very tedious to measure because someone must physically set up the device and drive along every road to be included. This process is extremely costly and time-consuming in large countries such

as DRC, with many roads in bad conditions drastically reducing travel speeds. Because of this limitation, IRI data is very scarce, particularly in developing countries with weaker infrastructure. Its scarcity is not just an issue for policy-making but also for establishing alternatives to measure road quality using machine learning. Training algorithms to evaluate road quality leveraging satellite data requires large datasets combining imagery and labels of road quality. Without large amounts of road quality data serving as ground truth, the algorithms cannot uncover underlying patterns in imagery and identify their relation to road quality. Having few resources providing reliable IRI measurements in developing countries is one of the reasons why there is little literature in this area.

In the two studies most related to our approach, the researchers use IRI data as the basis of their models (Cadamuro et al., 2018; Thegeya et al., 2022).

There exist a few potential alternatives to using IRI as a ground truth for road quality. One approach described in the literature is to use a less sophisticated technique for measuring road bumpiness. Brewer et al. (2021) use a cellphone application to record the movements of users' phones with the associated geolocation while they travel in vehicles over roads. Another possible approach to providing labels to machine learning-based methods is to identify and label road quality issues on image data manually. This alternative is unfeasible for satellite imagery, mainly because even with high-resolution satellite imagery, it is impossible to identify single potholes. Leduc and Assaf (2020) demonstrate the feasibility of manually labeling potholes on high-resolution street-level footage of GoPro cameras mounted on vehicles. Similar to IRI measurements, acquiring this kind of footage requires vehicles to drive along all road segments to be evaluated. This limitation is why this type of footage is incompatible with the research questions our study aims to answer.

## 2.2 Satellite Imagery

As the primary goal of this study is to evaluate road quality for hundreds of kilometers, we decided to base our approach on satellite imagery as it is a relatively easily accessible and versatile data source capturing road characteristics. In the studies using satellite data, the

specific imagery used can be characterized by different parameters. The most important is the resolution, defined as the distance on the ground corresponding to the height or width of one pixel in the image. Analyzing the literature, we found large variance in resolution used in the three studies most closely related to our project.

One would expect that this significant difference in resolution manifests in the prediction performance of the models using this data, a hypothesis that we will have a closer look at in the next section.

Besides the image resolution, we also found different data sources used in these studies. Most allow accessing satellite imagery from a specifiable satellite on a selectable date, thereby ensuring consistency between images that the algorithms are trained on.

Another essential aspect of combining satellite imagery with road-quality data to train machine learning models is how to split the imagery into patches that capture roads and are of a consistent size that a computer vision algorithm can process. This requires that the size of all patches is constant. Cadamuro et al. (2018) and Thegeya et al. (2022) trace along the roads and create a box over the center of the road. In the third study (Brewer et al., 2021), an algorithm tracks the path traveled by the application users and zooms to the coordinates of the drivers' locations. In this context, there are two main points to highlight. First, following the exact directions of the road by adjusting the angle of the bounding boxes cutting out the road segments results in distortions on a pixel level when feeding the imagery as rectangles to the neural networks. Therefore, all bounding boxes must be aligned with the horizontal and vertical axis. Second, the dimensions of the boxes play an important role not just because they have to be resized to the dimensions the algorithms require but also to exclude surrounding areas, potentially confusing the model. A CNN generally learns with enough training data which part of an image is relevant for correct classification. Correlational patterns like roads in urban areas being, on average, of higher quality may result in the network basing its predictions on the presence of houses rather than the characteristics of roads. By adjusting the image size, the surrounding captured on each tile can be limited. Researchers successfully used images of size 224x224, and also 64x64 that were then resized to 224x224 (Cadamuro et al., 2018).

| Paper | Year | Images | Labels | Resolution (m/px) | Total Images |
|---|---|---|---|---|---|
| (Cadamuro et al., 2018) | 2018 | satellite | IRI / 10 m | 0.5 | ∼115k est. |
| (Thegeya et al., 2022) | 2022 | satellite | IRI / 100m | 10 | 94,000 |
| (Brewer et al., 2021) | 2021 | satelite and Google Maps | app data | 0.3 | 53,686 |
| Our study | 2023 | Google Earth | IRI / 100m | 0.6 | 10,081 (256px)<br>20,876 (64 px) |

Table 2.1: Table comparing data quality and quantity.

**Other Inputs**

Satellite imagery captures characteristics of roads that allow for road quality assessment. Other factors varying with geolocation not captured by imagery are also expected to be important in predicting road quality and can therefore be included in the models. For example, tabular data such as average daily temperature, precipitation, land gradient, and the local population allow contextualizing the imagery and thereby increasing the model's performance. For Thegeya et al. (2022), this effect was very large and improved the performance of a binary classifier from 61% accuracy to 75%.

## 2.3 Methodology and Results

As previously elaborated, the IRI is a continuous unit of measurement and standard for quantifying road quality. Minor differences between IRI values are hard to interpret and not helpful in policymaking, where the main focus lies on distinguishing good from bad roads. Most studies, therefore, split the IRI's range into multiple classes and train classification rather than regression algorithms. Since there are only some vague guidelines and no clear standard for what IRI value categorizes a road as good or bad, it is difficult to capture the decision process underlying the exact cutoffs used in the literature. Further, different studies also use a different number of classes in the range of 2 to 5.

Neural Networks and particularly approaches based on Convolutional Neural Networks have become the standard in the Computer Vision literature. The architecture used in

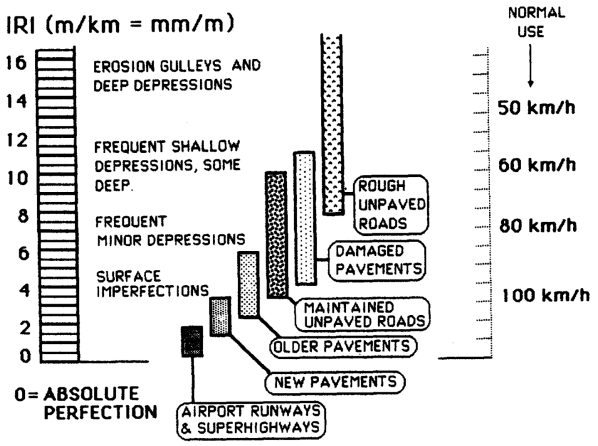



Figure 2.1: Some ranges for the IRI scale (Sayers et al., 1986).

Cadamuro et al. (2018) and Thegeya et al. (2022) are AlexNet, SqueezeNet, and VGG.[1] We can see that Cadamuro et al. (2018) achieves higher accuracy than Thegeya et al. (2022), part of which likely can be attributed to the higher image resolution. It is important to note that these results cannot be compared as accuracy as a performance metric highly depends on the class distribution.

| Paper | Cat. | Results | Best Model |
|---|---|---|---|
| Cadamuro et al. (2018) | 2 / 5 | 88% / 73% | VGG-11 and SqueezeNet |
| Thegeya et al. (2022) | 2 / 4 | 60% / 39% (87% / 75% w tab. data) | VGG-11 and SqueezeNet |
| Brewer et al. (2021) | 3 | 80% | InceptionResNetV2 |

Table 2.2: Table comparing results from the most relevant studies.

Besides the points already mentioned, the main takeaways from the literature review are the following: Firstly, we must ensure that labels and satellite imagery match up time-wise, as it is important to account for the fact that roads change over time. Secondly, there is reason to believe that increasing the number of training examples will not improve performance (Thegeya et al., 2022), but because in this case the authors used 10 m/px resolution, it may be that there was limited information to be learned from these images in the first place. Our goal should be to obtain a number of images on the order of the number

[1]SqueezeNet is designed to perform as well as AlexNet but with fewer parameters. VGG stands for Visual Geometry Group and improves performance in cases with lots of images.

used in Cadamuro et al. (2018) which achieved the best results reported in the literature and has an estimated 115 thousand image-IRI pairs.[2]

[2]We estimated this number based on the length of road segments, the image resolution and the image size reported in the paper.

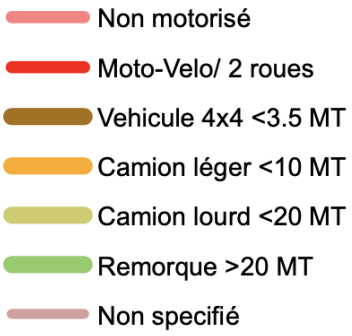
# 3. Data Sources and Preprocessing

## 3.1 Data Sources

As previously described, recording IRI data is very time and labor-intensive. Further, the region of interest for our study suffers from continuous conflicts, making it very hard to record the IRI data, which is precisely why we are undertaking this project. Unfortunately, the lack of IRI data also implies that no data available would be granular enough to serve as a label for training our model in the DRC. Fortunately, we were granted access by the World Bank to a data set of 112,901 IRI measurements for Liberia. Therefore, we decided to train our road quality assessment model in Liberia and then apply it in the Sud Kivu region of the DRC. Liberia shares a few climate and geographic characteristics that make this an avenue worth pursuing. First, Liberia is located on the same continent but around 6 degrees above the equator while Sud-Kivu is located around 3 degrees below the equator. Liberia has a tropical climate and Sud-Kivu has tropical and subtropical climates. Liberia is a little warmer, with annual temperature fluctuations ranging from 18-32 °C, while in Sud-Kivu from 14-28 °C (World Bank Climate Change Portal: DRC; World Bank Climate Change Portal: Liberia). Sud-Kivu receives significant rainfall (one of the main challenges facing people building infrastructure), on the order of 1600 mm annually. But it turns out Liberia is even wetter, with interior regions getting on the order of 2000 mm precipitation annually and coastal areas even 2500 mm on average. Besides many characteristics in common, there are also stark geographic differences, which must be considered. Liberia is a coastal country, while Eastern DRC lies in the continent's center and is characterized by a mountainous landscape. Further, the rainy seasons of both countries are not at the same period of the year. While the rainy season of the DRC lasts from September to May, that

of Liberia lasts from May to October.

(a) The full map.




(b) The road categories, from non-motorized to trailers weighing more than 20 MT.


(c) A close up of some of the categorized roads from the map.

Figure 3.1: One of the road type maps of Sud-Kivu made for humanitarian organizations.

### 3.1.1 Road Type Maps

The World Food Program created a series of maps for humanitarian organizations to navigate different parts of the DRC. These maps contain information about the vehicles that can be used on certain roads. We also found similar maps for Liberia, although the categories used for the road classification are different but follow the same logic. The maps are too large to be displayed here, but snapshots of a sample map are shown in Fig. (3.1).

### 3.1.2 Liberia IRI Data

The IRI data for Liberia consists of 112,901 IRI measures taken from April 1, 2016, until July 31, 2016. Each value corresponds to a road segment of around 100 meters documented in m/km.

### 3.1.3 Google Earth Tiles

High-resolution satellite data at the 30-50 cm level is expensive, so we looked for alternative data sources. We found an application called Google Earth Images Downloader[1] that queries Google Earth for satellite image tiles in 256x256 resolution. To download these images, one

[1]More information here: https://www.allmapsoft.com/geid/

must specify the four coordinates to define the rectangle of the area to be downloaded, the zoom level, and a date. The program only takes images as close as possible before the specified date. Because of cloud cover, brightness issues, and limitations on the range of satellite coverage, the map visible on Google Earth is a patchwork of images from different satellites taken from different angles on different days. This means that the quality and characteristics of images vary, which may present a challenge for our algorithm. Given enough data we expect the algorithm will learn to filter out these variations. Following Cadamuro et al. (2018) and Thegeya et al. (2022), we only included imagery taken within one year before to one year after the window in which the IRI values were measured. Unfortunately, there was only minimal imagery coverage at the highest resolution available during that period, so we decided to download the data at zoom level 18 (60 cm/px)(Stefanakis, 2017). For Sud-Kivu we also downloaded the tiles at the same resolution but for the date corresponding to the years in which the road-type maps were issued.

### 3.1.4 Shape Files

For the DRC we were able to find a shapefile from OpenStreetMaps, an open-source maps initiative.[2] This file contains 636,790 roads in the DRC including those in the Eastern region.

## 3.2 Preprocessing

Working with satellite imagery implies handling large quantities of data. One preprocessing step taking fractions of a second on a single image accumulates quickly to hours or even days for the entire dataset. It is therefore crucial to exclude data not containing the desired information as soon as possible in the preprocessing pipeline. With the program we used to download the satellite data, it is only possible to define the boundaries of a rectangle to download all images within. We, therefore, had to download over 3 million tiles covering Liberia and Sud-Kivu. Each image has a size of 256 x 256 pixels. For the road quality assessment model, we only needed images showing roads for which we had road quality data

[2]Downloaded from: http://download.geofabrik.de/africa.html

that were taken within one year of the quality evaluation date. Fortunately, the downloaded data contains, besides the pictures, metadata about the geolocation of each tile, its location within the rectangle of downloaded data, and the date the satellite image was taken. Using this information, we could exclude images irrelevant to our project based on metadata without having to interact with the files themselves. Following the approach of Cadamuro et al. (2018), to train our model we included imagery taken between one year before and one year after the quality assessment.

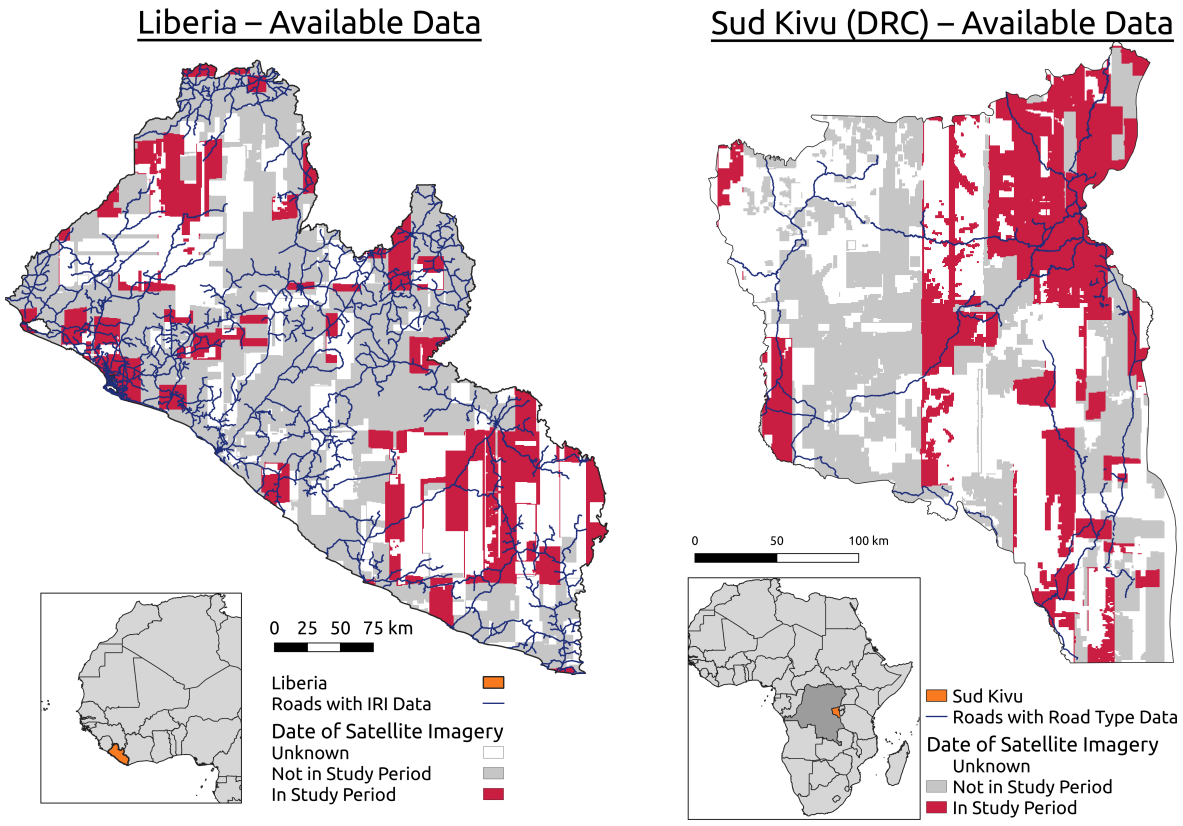



Figure 3.2: Map of downloaded satellite imagery for Liberia, categorized by date. The roads for which we have IRI data are shown in red.

As the IRI values in Liberia were recorded between April and July 2016, we followed the literature by reducing our data set to images taken between at most one year before or after this period. For Sud-Kivu, we only included data taken in 2018, the year in which the map qualifying different road types was published by the World Food Program. Fig. (3.2) shows our study regions highlighting the areas where we had satellite data for the relevant

period of our study. This corresponds to a total of 50 thousand images out of the 3 million downloaded initially.

Besides the date, the second essential inclusion criterion was whether an image contained sections of roads for which we had labels. The lines of shapefiles symbolizing roads sometimes do not represent the roads' exact location. To account for possible misalignment between roads and shapefiles, we created a buffered version of the shapefile with a buffer size of 10 meters. Having the roads represented by a wider buffer rather than a single line increases the probability that the space defined by the shapefile includes the roads' actual position on the images. There is, however, a trade-off between widening the buffer to include the exact position of roads with high probability and including more surrounding irrelevant parts within the buffer. This trade-off is a classical precision-recall trade-off where widening the buffer corresponds to increasing recall on the cost of precision and vice versa. To identify tiles containing roads computationally efficiently, we first cut the buffered shapefile to the spatial extent of our downloaded rectangle of satellite imagery. Then, we converted the buffered shapefile to a raster dataset specifying the number of rows and columns to exactly match those of our image rectangle. All raster cells in positions representing tiles containing no road segments are assigned zero. Combining this raster with metadata about the location of image tiles within the downloaded rectangle allowed us to rapidly identify tiles capturing roads even for large quantities of images. In Liberia, 30 thousand images contained some of the 10 meters buffer around roads. In Sud-Kivu, 20 thousand. Including only roads for which we had imagery taken within our study period, we had a sample size of 10 thousand tiles for Liberia and 6 thousand for Sud-Kivu.

After identifying the relevant tiles for our study, we calculated the exact IRI value assigned to each tile in Liberia. Since each road segment for which we had one IRI value was about 100 meters long, one tile possibly contained multiple different-sized parts of road segments, each assigned to an IRI value. To best represent the quality of roads in an image, we calculated the mean of IRI values of captured road segments within an image weighted by their pixels share. Therefore, we created a raster of the buffered shapefile for each tile with the same dimensions as the image's pixel size (256 x 256) (see 3.4). The value of each

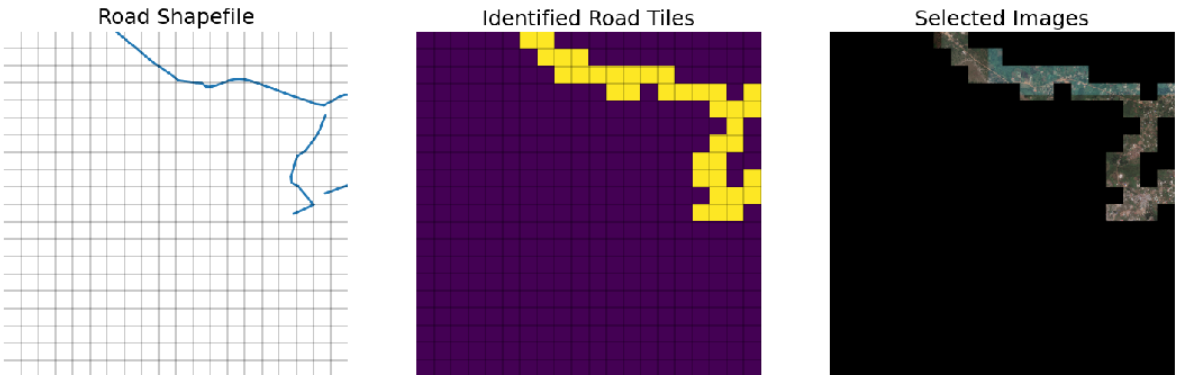



Figure 3.3: Identifying the images capturing relevant roads.

cell in the raster corresponded to the IRI value of the road segment at the same position. We then calculated the mean cell value for all non-zero-valued cells in the raster to get the IRI label for a specific tile.

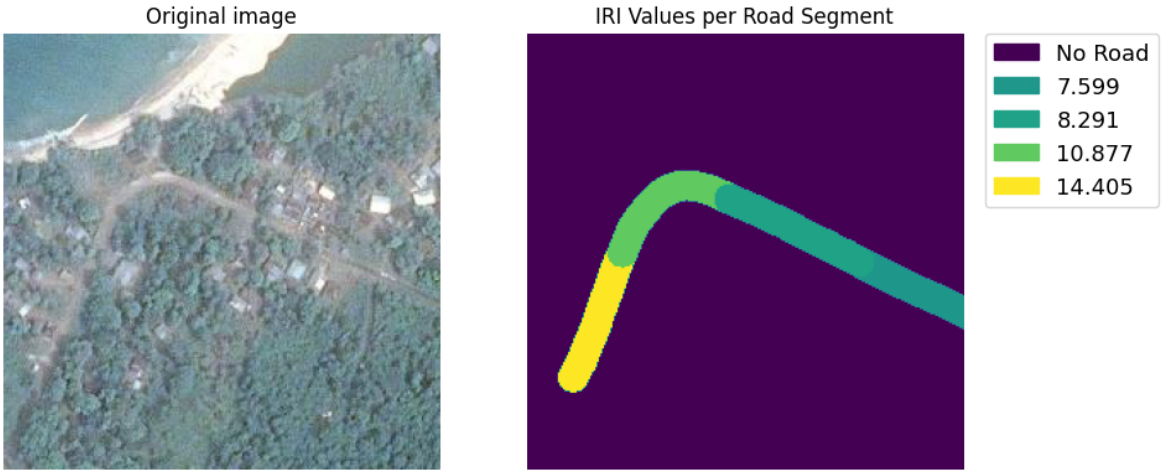



Figure 3.4: Assigning one IRI value per image

Previous literature found that considerable variation of road quality for road segments captured within one image or for road segments for which only one mean IRI value is available limits the CNNs abilities to learn characteristics associated with road quality. In our case, the IRI data is the mean IRI in meters per kilometer for segments of around 100 meters. While we cannot improve the granularity of our IRI data, one parameter we can manipulate is the image size which correlates with the length of road segments and the variation of road quality captured. As noted before, our images of 256 x 256 pixels

correspond to a height and width of about 150 meters. These images often capture parts of multiple road segments with different IRI values. Reducing image size might be helpful for the model to train. Still, since our IRI values are averages of 100-meter segments, it might also worsen the result because the model will only see portions of the sections based on which we assign the label to each tile. To understand how the spatial extent captured on each image influences the performance of our algorithm, we split our entire sample into images of 64x64 and reassigned labels using the raster-based approach previously described.

# 4. Methodology

For our study we are aiming to: (1) train road quality detectors that predict IRI categories in Liberia under conditions as similar as possible to the models in the literature for comparison purposes, (2) train another network using the same data but with intervals chosen specifically to detect roads of poor/bad quality, (3) provide holistic model performance metrics and quantify the uncertainty of our predictions, (4) apply our best performing model to roads in DRC and relate the results to road type maps published by the World Food Program.

## 4.1 Choice of CNN

We decided to use transfer learning for our classification task. Transfer learning is defined as using models pre-trained on large image classification datasets. Using models with pre-trained weights, training the last layers, and adjusting the number of output classes performs well on numerous image classification tasks.

We compare our models' performance against those in the literature using binary classification. The primary limitation in comparing results across road quality assessment literature is that authors use data from different countries and different classification thresholds. Further, the studies most relevant to our project use accuracy as the primary performance metric, a measure susceptible to changes in the data distribution. To make the comparison as accurate as possible, we decided to take thresholds reported in the two papers we want to compare our model with, namely an IRI value of 20 for Cadamuro et al. (2018), and 43.6% based on the distribution of IRI values as done by Thegeya et al. (2022). Besides these thresholds, we also report results for a binary classifier divided at the median IRI value of

our training set. To compare results, we used the best CNN-based model reported in both papers, VGG-11. In addition, we also work with EfficientNetV2, a CNN variation with good performance that trains 5x-11x faster than similar networks (Tan and Le, 2021). Reducing training time is crucial for such a data-intense task as it allows us to iterate and experiment with different settings, thereby being more efficient and increasing performance.

Besides our original 256x256 pixel image tiles down-scaled to 224x224 we also compare performance with 64x64 pixel tiles that are up-scaled to 224x224. Of course, reducing the image size also increases our final sample size from 10,081 to 20,876.

## 4.2 Choosing Appropriate Categories

Converting a continuous metric as IRI to a classification problem requires thoughtfully setting thresholds to split the distribution of values into classes. We assume policymakers are particularly interested in identifying roads in poor/bad conditions. Therefore, as our IRI threshold for poor/bad quality roads in our binary classifier we selected a value of 15 based on the scale presented in Fig. (2.1), because an IRI value above 15 selects for rough unpaved roads and worse. This threshold is more lenient than the one chosen by Thegeya et al. (2022), as they assume an IRI value above 5 is already an indicator of poor road quality. On the other hand, we are stricter than Cadamuro et al. (2018), who chose an IRI value of 20 as the breaking point.

The criteria for choosing our class intervals for the five-label classification task is conditioned on the previous literature and the inherent distribution of our data set. We use Cadamuro et al. (2018) instead of Thegeya et al. (2022) as a reference, provided the road physiognomy of Kenya resembles much more that of Liberia and DRC than the one in the Philippines. Recall the class intervals chosen by Cadamuro et al. (2018) read as follows: great (0-7), good (7-12), fair (12-15), poor (15-20), and bad (20+).

We adopted the same labels but implemented a series of modifications according to our data's distribution of the IRI values. Fig. (4.1) illustrates that the distribution clusters around the median at 14.4 and shows a positive skew. We've recalibrated the fair-good

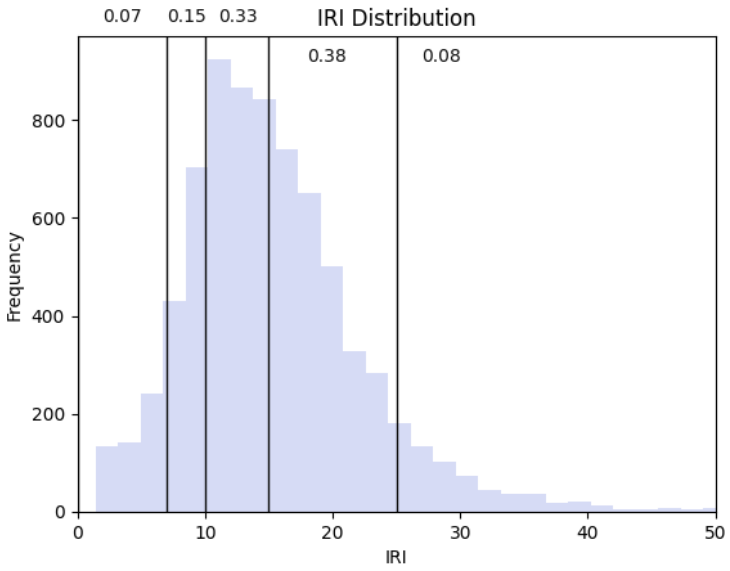



Figure 4.1: IRI distribution in Liberia with vertical lines indicating the class intervals and label indicating the relative distribution weight of each class

boundary from 12 to 10. This adjustment aims to encapsulate the natural cluster within the 10 to 15 range. Moreover, we've broadened the boundary between poor and bad by an increment of 5, balancing the proportion of measurements within the extreme classes. Through these adjustments, we've achieved that both extreme classes (great, bad) contain the same proportion of samples in our dataset.

# 5. Classification Models

In this chapter, we initially compare our results to the literature. We then fine-tune a 2-class and 5-class EfficientNetV2 with adjusted class boundaries. Finally, we calibrate the output probabilities and attempt to quantify uncertainty using conformal inference.

## 5.1 Binary Classification

**Comparison with Literature**

As the first step of our analysis, we aim to compare the predictive abilities of our model for images of the dimensions 64x64 and 256x256 to the results of Cadamuro et al. (2018) and Thegeya et al. (2022). We calculate the performance metrics reported by averaging them for the test set during the last three training epochs. As previously described, this comparison has to be taken with a grain of salt. Both papers use IRI data for other countries and report their results using accuracy as the primary evaluation metric. We follow this approach even though the different distributions of IRI values in each study make comparing model performance based on accuracy difficult. Accuracy is not a holistic metric of model performance because it is prone to be influenced by data distribution and class imbalance, so we also provide AUC. Moreover, accuracy does not distinguish between the different types of errors. The results are displayed in Table (5.1) and Table (5.2):

Our model achieves higher accuracy than Thegeya et al. (2022) but lower accuracy than Cadamuro et al. (2018). One reason for performing better than Thegeya et al. (2022) is the higher image resolution of our data set. Surprisingly, our 64x64 pixel images decreased the performance of our model.

| | | Accuracy | | AUC |
|---|---|---|---|---|
| Binary Threshold | Model | Cadamuro et al. (2018) | Us | Us |
| IRI = 20 | VGG-11 (64) | 0.90 | 0.62 | - |
| | VGG-11 (224) | 0.87 | 0.74 | 0.73 |
| | EfficientNetV2 (224) | - | 0.80 | 0.75 |
| median | VGG-11 (64) | - | 0.54 | 0.55 |
| | VGG-11 (224) | - | 0.64 | 0.67 |

Table 5.1: Accuracy results (and AUC) for our binary models using the same IRI split as those in Cadamuro et al. (2018). In our case, an IRI value of 20 gave us almost exactly an 80%/20% split. The median was at an IRI value of 14.32 (64) and 14.41 (256).

| | | Accuracy | | AUC |
|---|---|---|---|---|
| Minority Class Share | Model | Thegeya et al. (2022) | Us | Us |
| 43.6% | VGG-11 (64) | 0.6 | 0.53 | 0.52 |
| | VGG-11 (224) | 0.6 | 0.63 | 0.69 |

Table 5.2: Accuracy results (and AUC) for our binary models using the exact same split as those in Thegeya et al. (2022). We note that in their case 43.6% of the data was contained below an IRI value of 5, where as for our case it was for an IRI value of 13.40 (224) and 13.26 (64).

**Addressing Class Imbalance**

The main interest for policymakers in machine learning-based road quality assessment lies in the ability to identify a relatively small portion of roads that are in particularly bad condition and likely require reconstruction work. In such a problem setting, class imbalance naturally arises as not all prediction classes are the same size. There are two main techniques to address class imbalance. One is to up-sample (down-sample) the minority (majority) class in the training sample. Using this approach for binary classification, we sampled from the minority class by a factor equivalent to the proportion of the majority class and vice versa, thereby creating an artificially balanced training set. All the performance metrics throughout the paper are obtained after up-sampling the minority class. Besides eliminating class imbalance, we occasionally used this method to increase the training set sampling probability of the most relevant classes for our analysis. The second approach to deal with

class imbalance is to assign weights to the binary cross entropy loss function and thereby adjust the impact of wrongly classifying samples of each class on the total loss. By assigning a bigger weight to the minority class, the model will give it a larger relative importance when updating the weights during backpropagation.

### Performance Metrics with Adjusted Threshold

We trained the three common architectures used across the referenced papers (AlexNet, SqueezeNet, and VGG-11), and EfficientNetV2 outperformed all of them. That's why throughout this section and the rest of the paper, we will use the EfficientNetV2 architecture. Now, we train an EfficientNetV2 network using a threshold of 15 that, as stated in 4.2, is regarded as the breaking point between the cluster of poor and good roads. ROC curves indicate the algorithm's performance, providing meaningful information and facilitating comparison. Based on the ROC curve, we chose the best threshold to evaluate our predictions using the G-mean calculated from the training data, which helps us stabilize predictions for both classes. Figure 5.1 shows the ROC curve for the described binary classification splitting the IRI values at 15. The classification report with the optimal threshold is shown in Table (5.3). We note very balanced results for both classes.

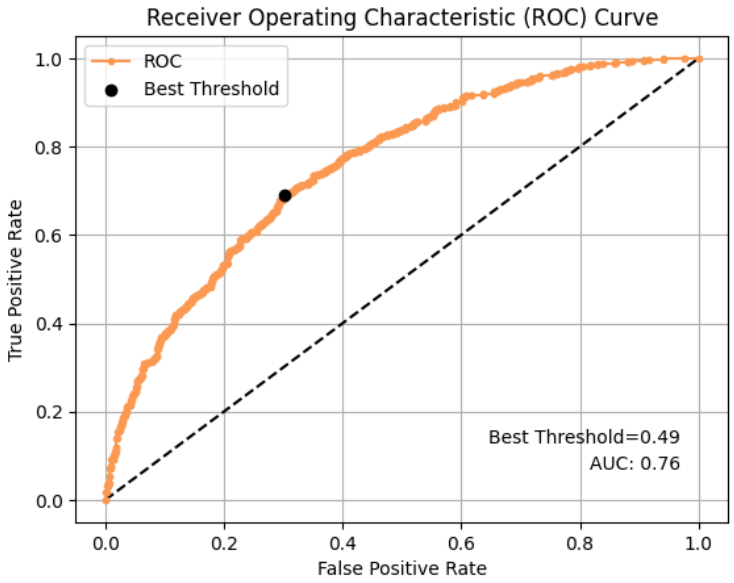



Figure 5.1: ROC Curve with reported AUC and marked in black the optimal threshold.

Table 5.3: Optimal Threshold Classification Report

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Class 0 | 0.72 | 0.69 | 0.70 | 792 |
| Class 1 | 0.67 | 0.69 | 0.68 | 706 |
| Accuracy | | 0.69 | | 1498 |

## 5.2 Multi-Class Classification

### Fine-Tuning Pre-Trained Model

We split our data set into three sets: training (80%), calibration (10%), and validation (10%). As explained in 4.2, the class intervals used in our project are driven by our data distribution and the previous literature. For our five-class classifications, we use the following IRI intervals: great (0-7), good (7-10), fair (10-15), poor (15-25), and bad (25+). We assumed that policymakers are primarily interested in identifying poor and bad-quality roads. Especially bad-quality roads account only for a small share of our data. To increase the percentage of bad roads (class 4) in our training sample, we resampled our original distribution $[0.07, 0.13, 0.33, 0.38, 0.08]$ by sampling from each class with a specified probability. After doing so, our new training set was distributed as follows: $[0.18, 0.11, 0.23, 0.31, 0.27]$.

We trained our EfficientNetV2 on this new training set for 25 epochs trying different combinations of batch size and learning rates for the Adam optimization algorithm. We finally took a batch size of 32 and a learning rate of 0.001 for training. We unfroze the final two layers for the model to have sufficient flexibility to fine-tune our data set. As EfficientNetV2 is prone to over-fitting, we introduced a dropout layer with a dropout of 0.3 and a batch normalization layer before the final linear layer. The batch normalization layer is also helpful to yield more calibrated probabilities (Guo et al., 2017). Data augmentation such as rotation, flipping, and trivial augment wide did not significantly impact performance.

### Performance Metrics Overview

Table 5.4 shows the classification report for the 5-class classification with the EfficientNetV2 model with the training set sampled as described above. The model is relatively successful in detecting instances of the 'poor' roads but struggles more in correctly identifying 'bad'

roads. Despite their lower representation in the simulated training set, we can observe that the recall for 'great' and 'fair' roads is higher than for 'poor' and 'bad' roads. It is worth noting that the 'great' roads are relatively well-classified despite forming a small class in the overall distribution. Further, 'bad' roads are not as well classified despite constituting the second largest category in the re-sampled training set.

Table 5.4: Classification Report for the 5-class model

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Great | 0.45 | 0.62 | 0.52 | 85 |
| Good | 0.31 | 0.20 | 0.24 | 152 |
| Fair | 0.45 | 0.56 | 0.50 | 362 |
| Poor | 0.56 | 0.48 | 0.52 | 431 |
| Bad | 0.43 | 0.40 | 0.42 | 94 |

Fig. (5.2), shows the OneVsRest confusion matrix for 'poor' and 'bad' roads. As the name suggests, OneVsRest transforms the multilabel into a binary problem by comparing one class against all others clustered together. Observing the OneVsRest results for our classes of interest allows a more precise interpretation of the model's performance through 2-dimensional confusion matrices.

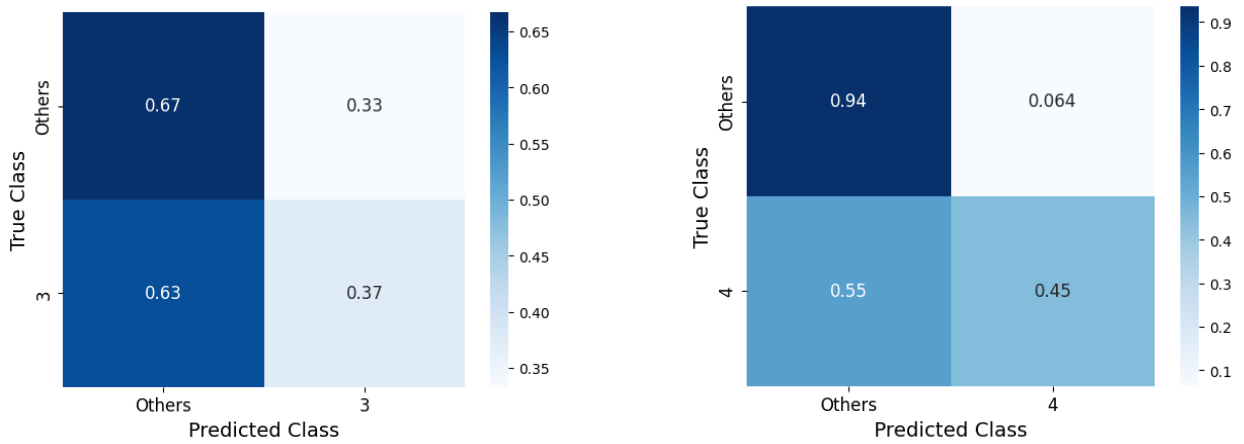



Figure 5.2: OneVsRest confusion matrix for the (a) 'poor' (class 3) and (b) 'bad' (class 4) road quality categories.

We thereby gain additional insight that the five-class classification report's precision, recall, and F1 scores might not explicitly reveal. One noteworthy observation for 'poor'

(class 3) roads is a large number of False Negatives, occasions where 'poor' roads were misclassified as not-'poor'. The model often misidentifies actual instances and performs poorly in this classification task. For the 'bad' category, the results are better. They show that the model almost doesn't misclassify not bad roads as being bad. The model generally shows a decent performance for predicting bad road quality.

## 5.3 Quantifying Uncertainty

None of the performance metrics reported so far allow us to quantify the uncertainty of our model's predictions. This section consists of two parts: model calibration and construction of prediction sets that contain the true label with a certain level of confidence using conformal inference.

### Calibrating Predicted Probabilities

A calibrated model predicts certain classes on average relative to the probability of these classes occurring. For example, in a zero-one binary classification task, an overconfident model would predict for a given subset 90% of the time a one even though only 70% of the true labels of samples are actual ones. After passing the logits of our model through a softmax function, we noticed the most likely predicted class was leaning towards $0.9 - 0.95$ for most of the observations, way more than its actual occurrence in the dataset. This was confirmed by iterating over a barplot representing the ranked classes.

In Figure 5.3, we see a plan of the process applied to compute the final adjusted probabilities by first splitting the data set. We then train our model on the training data. With the weights from this step, we predict each observation in the calibration set (passing the logit outputs through an external soft-max to convert them to probabilities). Then we train a calibration model on the calibration set using as inputs the predicted probabilities of the training model and as labels the true class belonging to each observation of the calibration set. We will employ a widely used calibration method, Platt scaling, with logistic regression in the calibration model. In addition, we use our model fitted on the training set to predict the test observations and obtain the output probabilities post-softmax. The final step

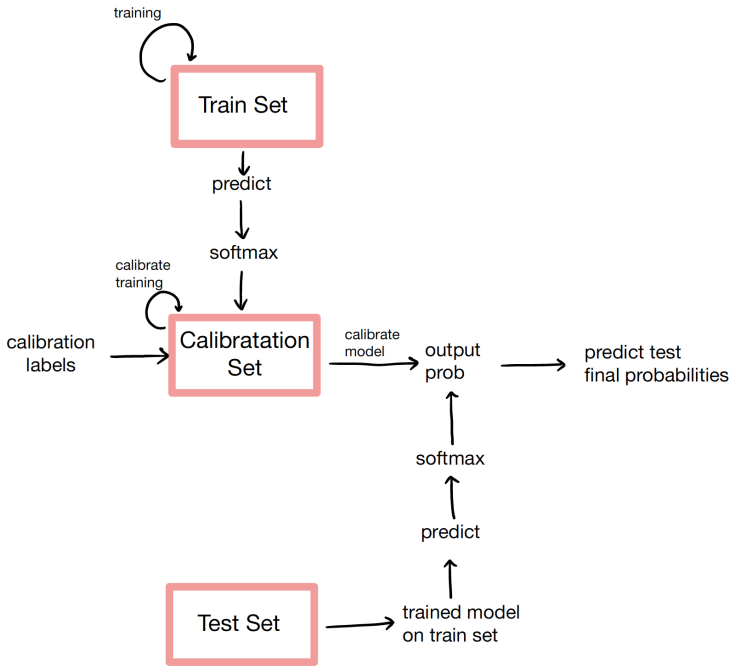
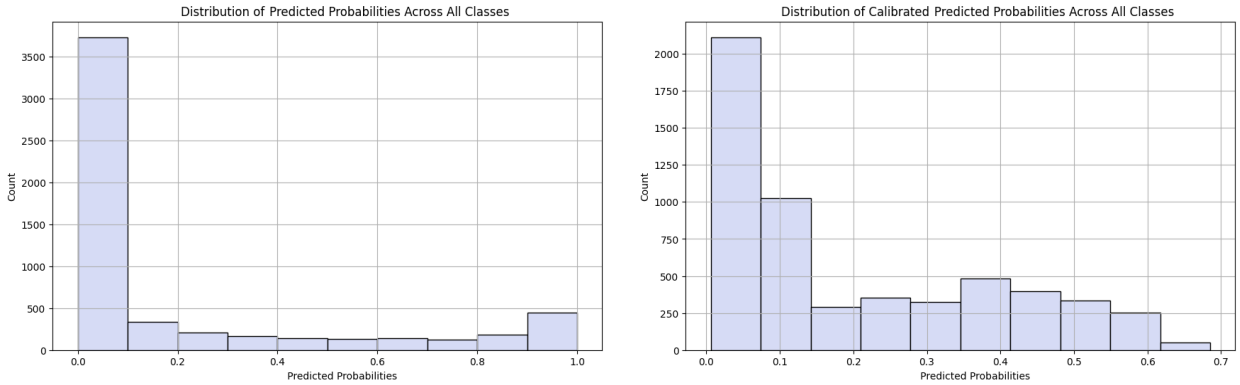



Figure 5.3: Data Set split and Model Calibration dynamics

involves fitting the calibration model on the predicted test set probabilities to get the final adjusted predictions.




(a) Before (b) After

Figure 5.4: Predicted class probability frequency before and after model calibration

In Figure 5.4, we see the predicted class probability frequency before and after the logistic regression calibrated the probabilistic outputs. Before the calibration step, the frequency distribution of the predicted class probabilities is biased towards the extremes, a clear indication that the model is not well calibrated and its predictions cannot be interpreted

as probabilities. After tuning the probabilities by the calibration model, the count range shrinks from a maximum value of around 3600 to 2100, meaning the predicted probabilities are no longer clustered in the extremes. Moreover, the upper bound of the predicted probabilities decreases from 1 to below 0.7, implying that the model is not producing overconfident predictions. Now that we improved the calibration of our model, we will be able to implement conformal inference and attain reasonably sized predictions on non-trivial prediction sets.

### Conformal Inference

Conformal inference provides a framework to quantify the uncertainty associated with a prediction independent of the underlying data distribution. It is a distribution-free method and enables us to construct valid prediction sets or intervals for any model, including deep learning architectures. The main objectives of the conformal procedure are to attain correct coverage (the probability that our prediction belongs to the prediction set) and having adaptive sets-prediction set size should reflect the difficulty of the classification task.

We will use a simple conformal procedure following Angelopoulos and Bates (2021) and assuming that non-exchangeability and symmetry hold as in Barber et al. (2022). The procedure is as follows:

1. Fit a model on the training set and use the test set[1] to obtain the softmax output probabilities,

2. Define as conformal score[2]:

$$E_i = \sum_{j=1}^{k} \hat{\pi}(x_i)_{(j)}$$

   where $\hat{\pi}(x_i)$ is the sorted softmax output and $k$ is the rank of the class. The conformal score $E_i$ is the total mass of softmax output until you reach the true class,

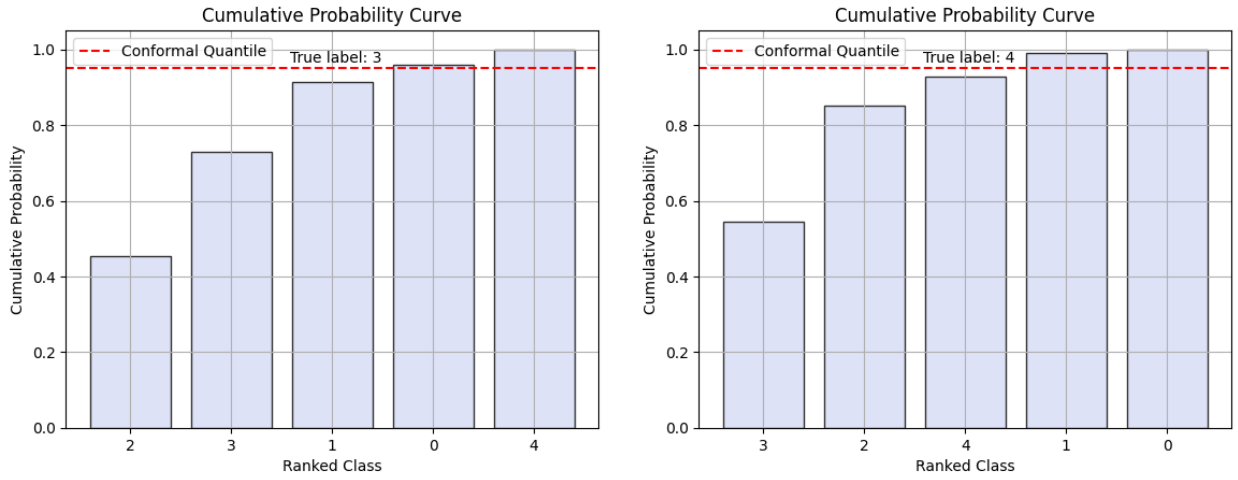
[1] We are not implementing split conformal inference, so we will not be using the calibration set for the conformal procedure

[2] nonconformity score measures the gap between the probability score for the correct class produced by the ideal classifier and the classification score produced by our model

3. Compute the 90th quantile of the conformal scores adjusted by a finite sample correction,

4. Select {the k most likely classes where $\sum_{j=1}^{k} \hat{\pi}(x_{n+1})_{(Y_{n+1})} \leq \hat{q}$}. As depicted in Fig. (5.5), we include in our prediction set the most likely classes up until the cumulative softmax probability is above the threshold $\hat{q}$.

If the total softmax mass exceeds $\hat{q}$, then we expect our true label to be within the predicted set around 90% of the time.




(a) Class 3 (b) Class 4

Figure 5.5: Prediction sets: For a) the true class is 3 and the prediction set is conformed by $2, 3, 1$ and in b) the true class is 4 and the prediction set is $3, 2, 4$

Figure 5.5 shows the ranked cumulative softmax predicted probabilities for our test set and the true label. For a), the true class is 3, and the prediction set is conformed by $2, 3, 1$, while in b), the true class is 4, and the prediction set is $3, 2, 4$. We need to distinguish between marginal and conditional coverage to affirm that the prediction set contains the true label with 90% confidence. Marginal coverage is the overall coverage among all groups, whereas conditional coverage is a stronger assumption that enforces the coverage to be at least 90% for every class.

As depicted in Table 5.5, the conditional coverage for class 3 is 0.97, meaning the model is particularly adept at producing prediction intervals that encompass the true class label

Table 5.5: Conditional Coverage per Class

| Class | Conditional Coverage |
|---|---|
| 0 | 0.68 |
| 1 | 0.697 |
| 2 | 1 |
| 3 | 0.9675 |
| 4 | 0.7234 |

for this class. However, for class 4, the conditional coverage dips to 0.72, suggesting the prediction intervals may not be as reliable for capturing the 'bad' road tiles. This lower conditional coverage for class 4 aligns with our previous observations from the precision and recall analysis (Table (5.4)): the recall for class 4 was relatively low, indicating that the model struggled to identify all relevant instances of the 'bad' road class.

# 6. Classifying Road Type

This project aims to analyze the capabilities of satellite-based approaches for road quality assessment in DRC. It is motivated by the fact that available road quality data in Eastern DRC is scarce and almost impossible to acquire. While the primary motivation for this project, the absence of available data is also its most significant hurdle. Having no granular data labels for DRC means it is impossible to train a computer vision model on satellite imagery of roads in DRC, and there is no other choice than to train the model in a different location with enough data available. Following this approach, the main question is how to ensure correct predictions while having a domain shift in the imagery. Can a model trained in Liberia correctly predict road quality in DRC? We try to answer this question by leveraging the available data sources for both countries. The maps provided by the World Food Program distinguish between roads that can be used by either light or heavy (all) vehicles. Assuming that this distinction correlates with the vertical displacement of a vehicle, in Fig. (3.1) we compare the distribution of IRI classes for road types passable by light and heavy vehicles. We do so for three different settings. Twice on the test set for Liberia, once with classes assigned using the actual IRI values and once using the predictions of our best-performing model. The third plot then shows the same distributions based on IRI classes predicted from satellite imagery of roads in DRC.

While the plots for Liberia do not seem to provide evidence that the road types on the maps are correlated with road quality, the plots for Congo paint a different picture. For Congo, there seems to be a clear relation between road type and road quality, as the share of bad and poor roads is higher for light vehicle roads, while at the same time, the percentage of good and fair roads is lower. This is surprising as our model was only trained with

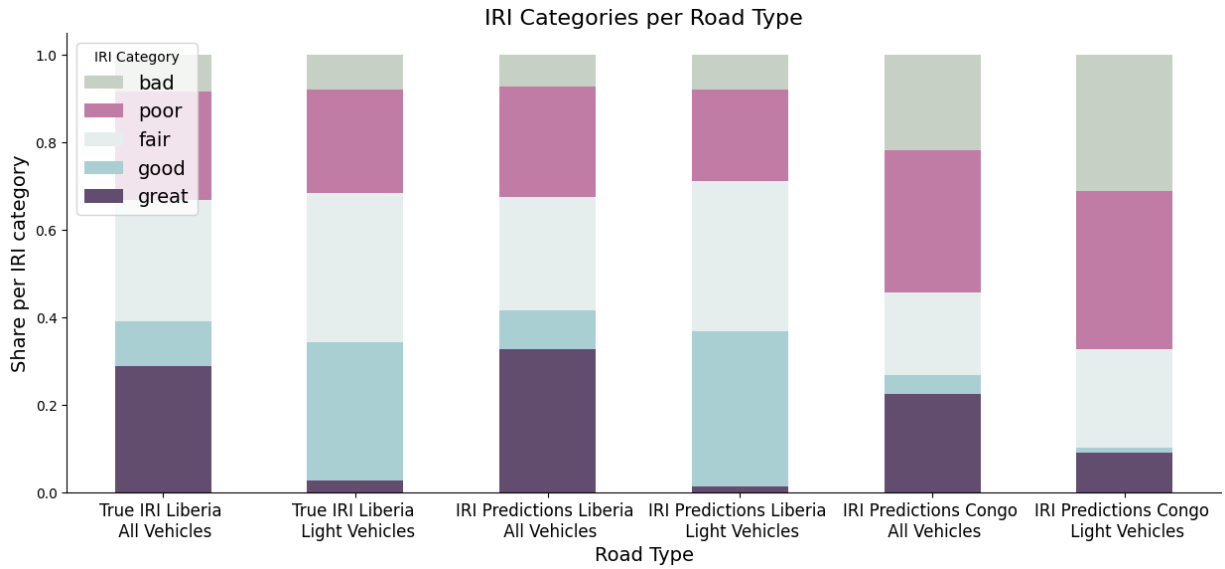



Figure 6.1: The distribution of road type classes in Liberia as well as the predictions for Liberia and Sud-Kivu.

labels and imagery from Liberia, where we cannot observe the same relationship for both the actual IRI values and the IRI predictions. What we can observe in all three settings is a stark reduction in good-quality roads between the all-vehicle and light-vehicle types. We expect this effect to be driven mainly by the pavement type, as it is easier for the model to distinguish paved from unpaved roads, which aligns with the literature (Brewer et al., 2021).


Figure 6.2: A set of five randomly generated images from each of the five classes for road tiles in Sud-Kivu.

# 7. Conclusion

In this project, we aimed to leverage satellite imagery to successfully identify bad-quality roads in Eastern DRC. We based our analysis on two classification scenarios by training a 2-class and a 5-class classifier. Due to the lack of road quality data in DRC, we trained both models on data from Liberia. We then evaluated their performance and found that both performed reasonably well compared to results achieved in similar studies. We used road-type maps as intermediaries to establish a connection between the models trained in Liberia and road quality in DRC. Even though it is impossible to quantify the performance of our model in DRC, its road quality predictions for different road types match our expectations. Further, manual inspection of images of each prediction category and proportions of predicted classes suggest these are reasonable results for roads in this region.

While we achieve some success in our models, we note several limitations worth elaborating. To take stock, one of our models performs slightly better than that of Thegeya et al. (2022), which uses 10 m/px resolution imagery. We believe predicting road quality with an image quality of 60 cm/px is a highly ambitious task. Any cracks or potholes will be completely invisible. Further, the concept of trying to measure bumpiness, caused mainly by the vertical displacement of the road, with pictures from above, precisely where it is most difficult to discern, presents quite a challenge. Distinguishing between paved and unpaved roads, as seen in the latter part of the study and already documented in the literature (Brewer et al., 2021), seems very well suited for this resolution level. But higher resolution imagery would probably be required to achieve high accuracy in this task. We would be quite curious to see how our algorithms would perform, and how this performance would change at different levels.

Moving to our particular experiments, we believe that our image collection technique (diverse images from different seasons and different satellites) could have affected the performance of our models. Given more data, our model could have learned more, but it would have undoubtedly helped us to have more standardized images from a single source. Despite the limitations mentioned here, we see promise in the approach of using satellite data to measure road quality; we expect that as quality of satellite imagery and the abilities machine learning algorithms continue to improve, so will their accuracy in detecting decaying roads for regions without any data to train on.

# Bibliography

A. N. Angelopoulos and S. Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *CoRR*, abs/2107.07511, 2021. URL https://arxiv.org/abs/2107.07511.

R. Barber, E. Candes, A. Ramdas, and R. Tibshirani. Conformal prediction beyond exchangeability. 02 2022.

E. Brewer, J. Lin, P. Kemper, J. Hennin, and D. Runfola. Predicting road quality using high resolution satellite imagery: A transfer learning approach. *PLOS ONE*, 16(7), 2021. doi: 10.1371/journal.pone.0253370.

G. Cadamuro, A. Muhebwa, and J. Taneja. Assigning a grade: Accurate measurement of road quality using satellite imagery. *CoRR*, abs/1812.01699, 2018. URL http://arxiv.org/abs/1812.01699.

L. F. Duffield. Vertisols and their implications for archeological research1. *American Anthropologist*, 72(5):1055–1062, 1970. doi: https://doi.org/10.1525/aa.1970.72.5.02a00040. URL https://anthrosource.onlinelibrary.wiley.com/doi/abs/10.1525/aa.1970.72.5.02a00040.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

E. Leduc and G. J. Assaf. Road visualization for smart city: Solution review with road quality qualification. *Internet of Things*, 12:100305, 2020. ISSN 2542-6605. doi: https://doi.org/10.1016/j.iot.2020.100305. URL https://www.sciencedirect.com/science/article/pii/S2542660520301372.

M. W. Sayers, T. D. Gillespie, and W. D. O. Patterson. 1986.

P. Schouten, J. Verweijen, J. Murairi, and S. K. Batundi. Paths of authority, roads of resistance: Ambiguous rural infrastructure and slippery stabilization in eastern dr congo. *Geoforum*, 133:217–227, 2022. ISSN 0016-7185. doi: https://doi.org/10.1016/j.geoforum.2021.09.017. URL https://www.sciencedirect.com/science/article/pii/S0016718521002633.

E. Stefanakis. Web mercator and raster tile maps: two cornerstones of online map service providers. *GEOMATICA*, 71:100–109, 06 2017. doi: 10.5623/cig2017-203.

M. Tan and Q. V. Le. Efficientnetv2: Smaller models and faster training. *CoRR*, abs/2104.00298, 2021. URL https://arxiv.org/abs/2104.00298.

A. Thegeya, T. Mitterling, A. Martinez Jr, J. A. Bulan, R. L. Durante, and J. Mag-atas. Application of machine learning algorithms on satellite imagery for road quality monitoring: An alternative approach to road quality surveys. *ADB Economics Working Papers Series*, Dec 2022. doi: 10.22617/wps220587-2.

World Bank Climate Change Portal: DRC. URL https://climateknowledgeportal.worldbank.org/country/congo-dem-rep/climate-data-historical.

World Bank Climate Change Portal: Liberia. URL https://climateknowledgeportal.worldbank.org/country/liberia/climate-data-historical.

World Bank Group - Press Release. World bank approves $750 million to support critical governance reforms, transport infrastructure, and digital connectivity in the democratic republic of congo, Jun 2022.

URL https://www.worldbank.org/en/news/press-release/2022/06/28/world-bank-approves-750-million-to-support-critical-governance-reforms-transport-infr