MAPPING ROAD PAVEMENT QUALITY FROM OPTICAL SATELLITE IMAGERY USING MACHINE LEARNING

BISRAT ARAYA GEBREEGZIABHER
August 2021

SUPERVISORS:
(M.Sc.) V. Venus
Prof. Dr. A.D. Nelson

MAPPING ROAD PAVEMENT QUALITY FROM OPTICAL SATELLITE IMAGERY USING MACHINE LEARNING

BISRAT ARAYA GEBREEGZIABHER
Enschede, The Netherlands, August 2021

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Spatial Engineering

SUPERVISORS:
(M.Sc.) V. Venus
Prof. Dr. A.D. Nelson

THESIS ASSESSMENT BOARD:
Dr. Ir. C.A.J.M. de Bie (Chair)
Prof. M. Zuidgeest (External Examiner, University of Cape Town)
Prof. Dr. A.K., Skidmore

DISCLAIMER

This document describes work undertaken as part of a programme of study at the Faculty of Geo-Information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Faculty.

ABSTRACT

Food loss occurring along the supply chain poses a major challenge in sustaining global food security. While agricultural production has improved significantly over the recent years, the facilities to manage this production have not kept up. This insufficiency results in post-harvest losses that occur after the harvesting of agricultural products. Post-harvest losses are prevalent issues in developing countries, thwarting the efficiency of agricultural food supply chains. Transportation has a substantial role in these losses since it is a vital link in the post-harvest chain. Particularly in developing regions, where road transport is the typical linkage, there is a decisive necessity to ensure the quality of transport facilitation. Ensuring quality in this sense means that the condition of roads has to be monitored, maintained, and rehabilitated. However, due to the lack of sufficient resources, these activities are not undertaken regularly. This aspect has resulted in the prevalence of poor-quality road that induces in-transit damages to perishable agricultural products such as tomatoes.

This study argues that spatial road quality information is a valuable tool in addressing these challenges. More importantly, enabling the convenient accessibility of this information is vital for resource strained regions such as Sub-Saharan Africa. Towards this goal, this research investigated the potential of mapping road pavement quality from freely accessible optical satellite imagery using machine learning methods. Accordingly, shallow and deep learning models were developed to extract road quality information from Sentinel-2 satellite imagery using reference data collected for a corridor running from Accra (Ghana) to Ouagadougou (Burkina Faso) with crowdsensing technology.

The results were encouraging in realizing the use of such a data source for convenient access to road pavement quality information. The deep learning model, i.e., U-Net, reported an F1-score of 37.93% and an IoU of 32.40%, outperforming the shallow ML alternative in the form of random forest. The inherent data imbalance prevents comparison with conventional segmentation task performance. The results, however, were comparable to analogous road extraction projects that utilized Sentinel-2 images. The study also contrasted the use of Sentinel-2 imagery to that of Planet imagery data to assess the relative potential of Sentinel-2 imagery in the task. The results showed that Sentinel-2 images were more suitable than the Planet ones in the pixel-wise classification of road pavement quality (RPQ).

Furthermore, a three-class RPQ classification model was presented to resolve the ambiguity surrounding severity classes. With an F1-score of 53.65% and an IoU of 46.03%, this model performed substantially better. Alternative to this approach, a flexible modeling paradigm based on probabilistic threshold moving was also explored. Aided with heuristics of precision-recall tradeoff and the probabilistic nature of ML model predictions, the study showed that predictions of the models could be molded to suit the utility desired.

KEYWORDS

Post-harvest losses, Road pavement quality (RPQ), Optical satellite imagery, Machine learning, Deep learning

ACKNOWLEDGEMENTS

First and foremost, I would like to thank God, for I have persevered and weathered through this difficult time with the strength and will he has granted me. I would also like to thank my parents, without whom I would not have had this privilege. Their constant support, belief, and encouragement kept me going at it each day.

My gratitude goes to Ujuizi Laboratories (Ujuizi Holdings B.V.) for providing the reference data used for this research. I am grateful for my supervisor, Venus Valentijn (MSc.). He has stuck with me throughout my highs and lows in the course of this thesis while giving me the space to explore my research. I am thankful for his practical insights that were helpful in many aspects of my work. Although we haven't had enough contact, I am also thankful for my second supervisor, Prof. Dr. Andy Nelson, for his understanding and oversight.

I would like to thank all the Spatial Engineering staff members. They have established a wonderful educational concept in Spatial Engineering, which I was inspired by. Their constant support during the program has got me to this point. I also extend my gratitude to ITC and the University of Twente, which offered me the opportunity of this program.

Last but not least, I am profoundly thankful for my brothers, family, and the friends I got to know during this program. It has been a year of tough challenges, and having these friends and families has helped me carry on strong.

"Onward to the cliff of everlasting paint of wonders,

Onward to the end."

Bisrat Araya Gebreegziabher
Enschede, 2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS AND ABBREVIATIONS

AI	Artificial Intelligence
AOI	Area-of-interest
API	Application Programming Interface
BIC	Bayesian Information Criterion
BOA	Bottom-of-atmosphere
CHEETAH	Chains of Horticultural Intelligence; towards Efficiency and Equity in Agro-Food Trade along the trans-African Highway
CNN	Convolutional Neural Network
DL	Deep Learning
FAO	Food and Agriculture Organization
FCN	Fully Convolutional Network
FL	Focal Loss
GMM	Gaussian Mixture Model
HSI	Hyperspectral Image
HLPE	The High Level Panel of Expert
HR	High Resolution
ICE	Iterated constrained endmember
IQL	Information Quality Level
IRI	International Roughness Index
LiDAR	Light Detection and Ranging
LULC	Land use and land cover
MESMA	Multiple endmember spectral mixture analysis
ML	Machine learning
MSI	Multispectral Image
NDBI	Normalized Difference Built-up Index
NDVI	Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
NICFI	Norway's International Climate and Forests Initiative
NN	Neural Network
PCA	Principal Component Analysis
PHL	Post-harvest loss
ReLU	Rectified Linear Unit
RF	Random Forest
RPQ	Road pavement quality
SAR	Synthetic aperture radar
SDGs	Sustainable Development Goals
SPICE	Sparsity-promoting iterated constrained endmember
SSA	Sub-Saharan Africa
SWIR	Short wave infrared
UAV	Unmanned aerial vehicle

1. INTRODUCTION

1.1. Background and Motivation

Realizing food security and preventing all forms of malnutrition are among the 2030 Sustainable Development Goals (SDGs) and the main concerns of the United Nations Decade of Action program. In that regard, several efforts have been made towards achieving these goals. Even though overall global progress has been observed, Africa and most developing countries still exhibit significant uncertainties in food security (FAO et al., 2020). In 2019, Africa logged a Prevalence of undernourishment rate of 19.1%, far from the proposed trajectory and an increase from the earlier records (FAO et al., 2020).

The current trend in the works to tackle this issue is focused on reducing food waste and loss. The ironic truth is that even though there is enough food being produced globally for everyone, one person in nine suffers chronic hunger (FAO, 2018). This aspect highlights that a significant amount, more specifically, one-third of food produced globally, is wasted (FAO, 2011b). With a potential cascading effect of economic losses across the food value chain, and increasing prices for consumers, these losses impede food accessibility to vulnerable groups, thereby affecting their food security (FAO, 2017). Reducing food losses and waste can increase food availability and reinforce food security by ensuring an efficient food value chain from agricultural production to the consumer (FAO, 2017; van Berkum et al., 2018).

Recently, food losses that occur after the harvesting of agricultural products, i.e., post-harvest losses (PHLs), have become hot topics of discourse, especially in Sub-Saharan Africa (SSA) (Sheahan & Barrett, 2017). The first World Food Conference of 1974, which aimed to half the 15% PHL estimate of that time by 1985, marks PHL's early attention (Parfitt et al., 2010). Since then, several methods and technologies have been employed in Africa to respond to PHLs, most of which were of insufficient success and adoption (World Bank, 2011). PHLs remain an ever-present problem, particularly in SSA (Affognon et al., 2015).

One systemic contributory factor to PHLs is that even though global agricultural production capacity is increasing, food consumption habits in developing countries are also simultaneously changing (Kearney, 2010). A considerable effect of these changes is the increased attention to food quality and supply chain traceability (Bollen et al., 2006). This transformation is mainly because of the public's increasing concern about the accessibility and safety of agricultural food products (Hastuti, 2008). Moreover, despite advancing agricultural technology, climate change is growingly straining food production in many food-insecure areas, which further emphasizes the need to reduce these food losses (Bellù, 2017)

The largest PHLs exhibited in SSA are in fruits, vegetables, root crops, and tuber crops, which is primarily attributed to their perishable nature and the lack of suitable post-harvest infrastructure in the region (Affognon et al., 2015). Estimates of worldwide post-harvest losses range from 20-60% for fruits, vegetables, roots, and tubers and 20-30% for cereals and legumes (FAO, 2011a). Furthermore, qualitative losses can occur at several stages of the value chain resulting in these products being sold at reduced prices and virtually incurring economic value losses (Kitinoja & AlHassan, 2012).

These findings make the argument for a need to reduce PHLs primarily to improve food security in most affected areas such as SSA. Studies that aim to reduce PHLs open up the possibility of improving people's economic state throughout the food supply chain. Nevertheless, in contrast to the attention given to lessening farm-level losses, technologies developed to address off-farm PHLs are limited (Affognon et al., 2015).

As the means by which agricultural products are moved from farms to markets and consumers, transport has an essential role in post-harvest agricultural linkage (Tunde & Adeniyi, 2012). Among the various surface transport systems, road transport (trucking) is the dominant mode in most countries, and particularly in developing regions, attributed to its intrinsic flexibility, reliability, and relative planning simplicity (Londoño-Kent, 2009; The World Bank, 2020). However, developing countries remain hindered by inadequate surface transport systems that serve as the veins of inland food transport from “farm to fork” (Londoño-Kent, 2009; World Economic Forum, 2017). Primarily, poor quality of roads, largely evident in most African countries, exacerbated by inappropriate transporting practices, develop adverse conditions for food transportation resulting in substantial losses (Kojo Arah et al., 2015). In-transit vibration and shock caused by defects on bad roads have been proven to inflict damage on most fruits (Fadiji et al., 2016; Fernando et al., 2019; Jarimopas et al., 2005; Van Zeebroeck et al., 2006; Wasala et al., 2015), vegetables (Chonhenchob et al., 2009; Pretorius & Steyn, 2019), and roots and tubers (Rees et al., 2001; Shiina et al., 2013) thereby resulting in significant PHLs. Therefore, ensuring adequate quality road routes for agricultural food transport plays a vital role in reducing PHLs, increasing the efficiency and sustainability of food value chains, and consequently improving food security, especially in developing countries. Eliciting from this motivation, the following subsections describe the conceptual framework, point of entry for intervention, related works, and identified gap for this study.

1.2. Conceptual Framework

There are various conceptualizations of PHLs as a result of divergences about timing (e.g., pre-harvest, harvest, post-harvest), scope (e.g., criteria for loss), and terminology (e.g., waste and loss) (Chaboud & Daviron, 2017). This study will adopt the HLPE (2014) definition of PHLs with a food security perspective as a loss criterion. Accordingly, food loss refers to a decrease, at all stages of the food chain before the consumer level, in quantity or quality (FAO, 2014), of food that was originally intended for human consumption, regardless of the cause (HLPE, 2014). Quantitative food loss refers to the reduction in the mass of food, and qualitative food loss refers to the decrease of quality attributes of food, i.e., reduction of nutritional value, economic value, food safety, and consumer appreciation (FAO, 2014). Furthermore, specifying to PHLs, HLPE (2014) identifies post-harvest as the stage between harvesting and processing, notably excluding the processing stage. Commercial or economic loss translates the various types of losses into economic and monetary terms (Grolleaud, 2002). Adhering to these definitions establishes the boundary of this study.

In their re-framing work on PHLs, Tröger et al. (2020) indicated that studies on PHLs could benefit from viewing the supply and value chain as a system of human activities. The causes of PHLs should not be ascribed to one stage or actor in the food value chain; they should instead be viewed as being interconnected at micro, meso, and macro levels (HLPE, 2014; S. K. Tröger, 2019) and spatial scales (S. K. Tröger, 2019). The three causal levels and their adaptation to the context of this study are described in Table 1.1. Qualitative and quantitive damages to harvested food products due to in-transit vibration, identified as the immediate cause of PHLs at the transport stage, are the micro-level causes. These vibration and shock damages can be attributable to several meso-causes at various stages and scales, as explained in Table 1.1. However, from the various meso-causes, a common denominational causal factor can be deducted as the lack of convenient information on road conditions and their implications. It can be argued that unawareness—regarding apparent road conditions—of the responsible actors and coordinating bodies relevant to the food supply chain results in the listed meso-level causes. Nelson et al. (Nelson et al., 2006) highlighted the importance of road quality information in cutting the costs imposed by poor road quality. A transporter aware of on-coming road quality conditions can better plan a travel route to reduce losses due to PHLs and vehicle damage. Traders and farmers can make informed decisions in procuring transportation options, i.e., choosing vibration resilient transport vehicles or less damaged routes. More importantly, reliable information on existing road quality will have crucial significance in the effective and efficient planning and prioritization of road rehabilitation and construction projects. Additionally, information on the existence of difficult road conditions can encourage the formation of closer (to farming site) market locations that can intercept food products with more minor damage. The overarching inadequacy of road infrastructure translates to a macro-level systemic challenge the gives rise to consequent emergent causes of PHLs.

Table 1.1 The three levels of causes of PHL taken from HLPE (2014) and with their adaptation to this study

Level of causes	**HLPE** (2014) **definition**	**Adaptation**
Micro-level causes	Those resulting from actions or non-actions of actors at the same stage of the food supply chain where the loss occurs	Vibration and shock due to poor roads causing damages at the transport stage that translate into PHLs
Meso-level causes	Those related to any stage or the whole of the food supply chain Arising from the organization and relation of actors across the chain	Inappropriate packing of agricultural food products and procurement of transportation options that lack consideration of road conditions by traders and sometimes by farmers (when there is no trader involved in between) Poor travel route planning by transporters unaware of existing road conditions Poorly spaced vehicle repair facilities Inadequate rehabilitation and construction of roads and difficulty in their prioritization
Macro-level causes	Those that can be explained by more systemic issues and favor the emergence of subsequent level causes	Lack of adequate road infrastructure to support the food supply chain

According to the above interpretation, a conceptual framework shown in Figure 1.1 was developed to describe the cause and effect relationships across the food supply chain. The supply chain actors (displayed in light gray boxes at the bottom row) were outlined with the Sub-Saharan country situation and perishable products in mind by aggregating different terminologies obtained from the works of Robinson & Kolavalli (2010) and Van Wesenbeeck et al. (2014). The dark gray boxes describe the various causes and effects across the supply chain categorized by their levels (row labels). The downward solid arrows show cause and effect relationships, while the dotted downward arrow from macro to meso level describes an emergence relationship from the former to the latter. It is used to illustrate that the need for information on road quality is an emergent behavior resulting from the inadequacy of road infrastructure. The state authority in the framework is used to represent an aggregate of governing entities that oversee and facilitate the whole food supply chain. Consequently, it is shown to be responsible for the rehabilitation and construction of roads. The two levels in the effects row show the direct effects, i.e., PHLs, and indirect effects of the overhead cause. The following subsection elaborates on the causal relationship of road quality and PHLs and establishes an argument for the meso level cause of inadequacy of conventional road condition information as per the defined conceptual framework.

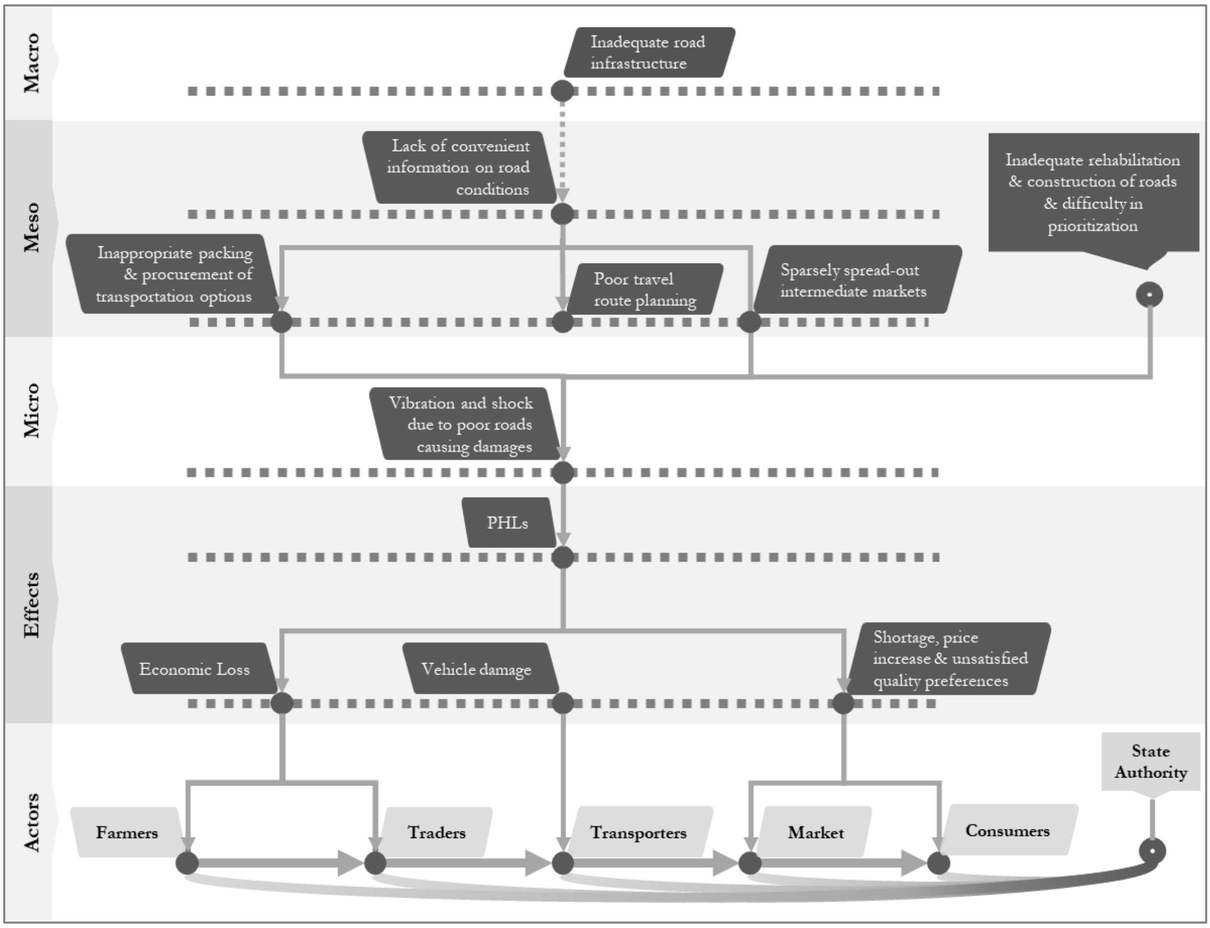


Figure 1.1 Conceptual diagram of the causes of PHLs in relation to transportation

1.2.1. Road quality information and PHLs

Road transport is often the viable option in developing countries to deliver perishable food products from farms to consumers as it offers shorter travel times and flexibility (Pretorius & Steyn, 2019). However, according to previous studies, the road imposes physical damage to sensitive food products such as fruits and vegetables due to in-transit vibration (Jarimopas et al., 2005). Besides the resulting loss in visual quality that deters consumers, researchers have also shown that physical damage also speeds up spoilage and loss of nutritional value, thereby collectively imposing considerable PHLs (Opara & Pathare, 2014). The trader and the farmer will hold the economic burden. At the same time, the market and consumer end will suffer the resulting supply insufficiencies, e.g., food shortage and price increase (see Figure 1.1). It is also important to note that transporters can incur costs from mechanical damages on their trucks due to poor road conditions. If the vehicles take a long time to fix, it will significantly delay delivery, risking spoilage of their products.

Soleimani and Ahmadi (2015) identified that road (surface or pavement) roughness is a critical factor in vibration-caused fruit damage. Physical damage to fruits and vegetables during transportation, regardless of their maturity before loading and packaging, is directly related to road roughness (Chonhenchob et al., 2009; Jarimopas et al., 2005). This aspect makes road roughness an important indicator relating road quality with PHLs. Sayers et al. (1986) define road roughness as the "variation in surface elevation that induces vibrations in traversing vehicles." Several indices have been used to quantify road roughness, among which the international roughness index (IRI) (Sayers et al., 1986) is the most used measure worldwide due to its versatility, practicality, and objectivity (S.-L. Chen et al., 2020).

Road surface distress, on the other hand, although not having a standard of measure like IRI, has recently been a common topic of interest in research and public applications related to road quality information. Surface distress generally describes road defects such as cracking, potholes, transverse and longitudinal deformation (e.g., rutting), and other miscellaneous ones (Wang, 2018). The taxonomy and explanation of the different types of road surface distress as described in the work of Paterson (1990) can be found in Appendix: Annex 1. These road defects are prevalent issues in African roads since their monitoring and rehabilitation are usually untimely. Surface distress information regarding the type, extent, and severity of the distress is valuable in scheduling maintenance activities since these defects have a deteriorating influence on the functionality of the road (Robinson et al., 1998). If timely maintenance (such as crack sealing and pothole patching) is not made at the early onset of visible defects, more costly measures (i.e., large-scale rehabilitation or reconstruction) might be required in due time (Robinson et al., 1998).

Several studies have analyzed the relationship between IRI and surface distress. Most of these investigations obtained high correlations (Rajendra Prasad et al., 2013), particularly with severe defects such as potholes (Mubaraki, 2016). However, it should be noted that different defects have different effects on IRI. Therefore, a reasonable conclusion from these works is that surface distress and roughness have a commonly causal relationship (Mubaraki, 2016). Surface distress is also known to cause difficulty in measuring road roughness due to the strong perturbation caused by the defects that distort calibration of modern vibration-based roughness measuring techniques (Wang, 2018). Moreover, information on road conditions regarding surface distress is often collected with semantics, such as the type and severity of the defect, that offers insight for decision making, even if it comes with the difficulties associated with visual assessments. This level of semantics is often challenging to achieve with roughness measurements such as IRI. This aspect makes pavement distress information important at both low and high levels of pavement management decision-making.

Road roughness and surface distress serve as essential measures in providing road quality information. However, it is vital to note that road quality information can vary significantly depending on available (or preferred) data collection methods and the desired utility of the information. In realizing the value of this information in road pavement management and transport facilitation, Paterson and Scullion (1990) formulated the concept of information quality levels (IQLs). They provide a solid foundation for establishing the amount of detail that individual data items must attain to support various levels of management tasks (Wang, 2018).

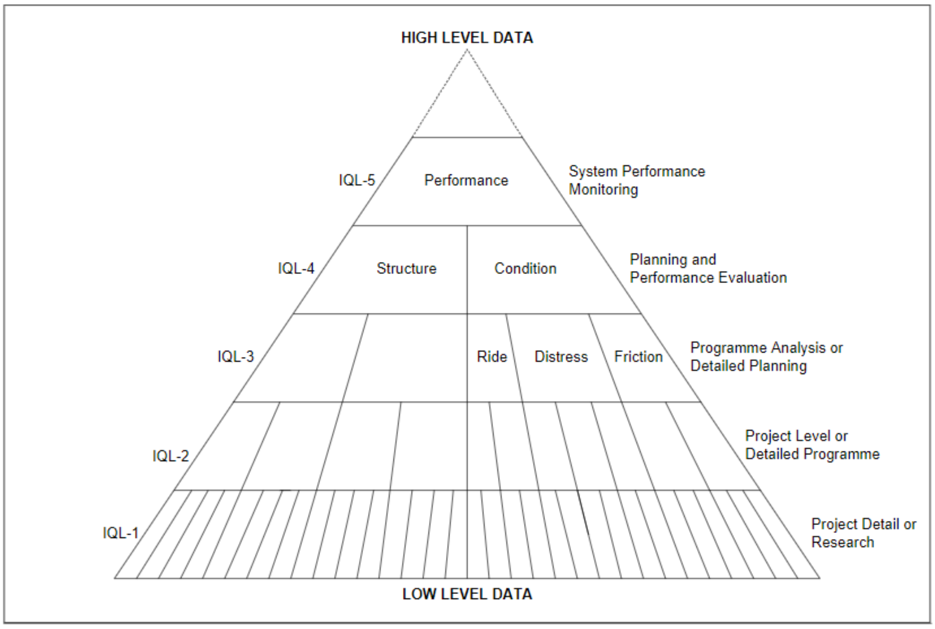


Figure 1.2 Information quality level (IQL) concept taken from Bennett and Paterson (2000)

The following description was summarized from the work of Bennett & Paterson (2000) in proposing the IQL framework. In this framework, information at low level (i.e., low-level data) representing comprehensive detail aggregates to progressively summative information at high levels of IQL (i.e., high-level data), as shown in Figure 1.2. At IQL-1, a thorough form of pavement condition information with more than 20 attributes collected from research, laboratory, theoretical, and electronic collection sources is represented, typically, to support project-level decision making. Close, in detail, to the previous level, IQL-2 represents reduced attributes obtained from engineering analyses. Bennet & Paterson (2000) define IQL-2 as having a simple level of detail, typically described by roughness, surface distress, and skid resistance—as relevant to road condition—appropriate for road network-level decision-making. At IQL-4, a summative attribute such as the description of road pavement condition in class values (i.e., good, fair, poor) or a categorical index (0-10) is presented for planning and management or in the context of low data collection capacity. Finally, IQL-5 represents vital performance indicators of road infrastructure obtain by combining road conditions with other measures (Bennett & Paterson, 2000).

An area of interest for this study would reasonably be the level IQL-4. This specificity closely relates to the linking nature of road quality information identified in the conceptual framework (see Figure 1.1 in Section 1.2). While lower IQLs serve a domain-specific purpose of detailed road monitoring, the higher levels at IQL-5 and beyond, as proposed by Bennett et al. (2000), relate to regional statistical indicators of road among those of other infrastructures. The middle point is where the two levels of authority and concern connect, and it is where, ideally, understandable and accessible information is preferred to a precision. Bennett et al. (2000) illude upon the need for such information to be easily understood without much technical background in the interest of high-level road management and the public. Users of the road, i.e., transporters and traders in the case of this study, can also be beneficiaries of such information, as mentioned previously.

Given that these users are less likely to find more value in detailed and high-accuracy road information, it is easily understandable that IQL III and IV would be the recommended levels of information for users' utility in planning their transportation. The value in the offering of road quality information for users lies in frequency and convenience. While reliability and resolution remain of fair importance, frequently updated and easily accessible information on road quality is of greater value for the users. More frequent collection of such information reduces the granularity of aggregation with which it is disseminated to higher-level decision-makers. These authorities are responsible for prioritization and procuring road rehabilitation/construction projects. The reduced granularity allows them to make better decisions with significantly less uncertainty. IQL-4 and perhaps IQL-3, where efficiently possible, are, therefore, the intersecting levels of information for the relevant stakeholders in the food supply chain.

Accordingly, this study adopts the term Road Pavement Quality (RPQ) from Ujuizi Laboratories (2018), which is an IQL-4 (arguably IQL-3) road condition information based on a categorical rating of road quality with three classes: good, bad, and very bad. It is used to evaluate the effects of typical road imperfections such as cracking, potholes, speed bumps, rough patches, bridge expansion joint, rumble strips, corrugated surfaces, sunken utility, etc., Ujuizi Laboratories, (2018). Although it involves identifying the type of road defects (distresses) and their severity in the data collection, it is fundamentally adopted in this study as a measure of the severity of surface distress. Section 4.2.1 discusses, in detail, the collection method of this reference data. Concluding on the complex dilemma of PHLs and poor road infrastructure, the following section formulates the problem as a 'wicked' problem and proposes an intervention based on the wicked problem framework (Georgiadou & Reckien, 2018).

1.3. The Wicked Problem of PHLs

Based on the wicked problem framework by Georgiadou & Reckien (2018), adapted from Hoppe (2010), a wicked problem can be characterized by uncertainty regarding facts, causes, and effects in one dimension and dissensus among stakeholders with respect to policy goals and values in another dimension. Starting with the uncertainty dimension, PHLs are characterized by a lack of consistent and clear knowledge regarding their occurrence, magnitude, causes, location (spatially and along the value chain), and their extent, leading to sub-optimal solutions and policy faults (Affognon et al., 2015). This aspect is particularly apparent in SSA (FAO, 2011b; Parfitt et al., 2010; Prusky, 2011). The multi-level cause and effect relationships explained in Section 1.2 emphasize the uncertainty of identifying a single cause and effect path for PHLs. An important intuition here is that, from a systems perspective, identifying the interrelated system of causes is the precursor to recognizing possible mitigation and priorities for action (HLPE, 2014).

Along the consensus dimension, developing scenarios in which each actor (agent) in the system is given autonomy to resolve the problem allows for examining alternative formulations of the problem and, thus, understand the existing dissensus. Table 1.2 summarizes these scenarios using four characteristics: perception, effects, accusation, and response. Perception describes the extent to which each isolated actor in the supply chain perceives the causes of PHLs. Effects indicate the resulting negative implications of PHLs to the corresponding actor. Accusation identifies the agent(s) that the respective actor would blame for the adverse effects of PHLs. Lastly, response describes the actions each would take had it been up to them to mitigate the effects of PHLs. This multi-perspective scenario evaluation illustrates the divergence in the conceptualization of the problem of PHLs and how that divergence provokes uncompromising responses from each actor, further consolidating the existing dissensus.

Table 1.2 Supply chain actors' perspectives and responses on PHLs and their causes

Actors	***Perception***	***Effect***	***Accuse***	***Response***
Farmers	- Lack information on causing factors	- Price loss, less frequent trader visits - local sellers could suffer direct PHLs	-Traders for low price rates	- Sell as quickly as possible (can lead to overpacking)
Traders	- Acknowledge that physical damage causes PHLs but lack awareness of the causes of physical damage	- PHLs resulting in economic loss - opportunity losses	-Farmers for poor quality products	- Overpack to account for the loss - Reduce visits to farmers - Decrease buying prices and increase selling prices
Transporter	- Acknowledge that physical damage causes PHLs but not the various causes of physical damage	- vehicle damage due to bad roads - economic loss if paid via sales	-State Authority	- Less frequent transportation trips - May take alternate routes to avoid bad roads
State Authority	- Acknowledge that poor infrastructure can indirectly cause PHLs	- strain on budget to maintain roads	- Transporters for overloaded trucks damaging roads - Traders for low quality and over-profiting	- set priorities for road rehabilitation & construction - establish rigorous regulations for quality control
Market & Consumers	- Relatively less knowledge on the causing factors	- Poor quality - Price increase - Shortage	-Traders for poor quality products	- opt for increased control of quality - resort to alternative sources, e.g., imported food products

Based on the above argumentations, it is reasonable to frame PHLs as a wicked problem. Therefore, towards structuring this problem as per the wicked problem framework of Georgiadou & Reckien (2018), this study navigates across the spatial knowledge dimension to decrease the uncertainty regarding the whereabouts of the causes of PHLs, particularly on the transport stage. Spatially locating the causative indicators can offer immense insight towards a comprehensive understanding and well-informed decision-making on PHLs. The agricultural supply chain as a logistical system has a better chance to make a holistic transition to efficiency if it can recognize where it is inefficient. Following the systemic approach described in the conceptual framework, three levels of causes to PHLs were identified. To answer the question of "where" in this context requires a balance between the degree of abstraction to allow entry points for interventions and the level of detail to link with the ground reality. Identifying causes at higher abstraction, e.g., recognizing a low-grade corridor connecting farm sites to urban markets, conceals the entry point for intervention, behavioral changes, and prioritization of investment. On the other hand, a higher level of detail in mapping causes, i.e., micro-level causes such as physical damages, is open to multiple interpretations such that a clear roadmap for action cannot be developed from such information. This gap calls for a bridging intervention between the macro and micro (Bergström & Dekker, 2014), which ideally lies in the meso level. However, it is essential to recognize the necessity of a mediating element as an intervention towards establishing trust and relative consensus. Reducing information asymmetry across the food supply chain has been recognized to have a vital role in establishing trust among actors and ensuring economic fairness (Minarelli et al., 2002). Therefore, it is essential, intervening through the provision of information on road conditions at the meso-level in this case, that this information is provided at the convenience of all relevant actors. Laubis et al. (2019) identified the facilitative value of frequent and timely available road condition information in the relation between users and authorities that manage roads. As per the discussion in the previous subsection, Road Pavement Quality (RPQ), serving as the adopted indicator for road condition information, fits this role suitably.

Mapping RPQs can offer decision-supporting information in the systemic transition of food supply chains towards sustainability. Primarily, as the overlooking authority, the state can use such information to draw out points of action and strategies that address PHLs relating to food transport with due consideration of other areas of concern. The essential utility aspect comes in distinguishing the need for action, thereby helping the act of prioritization of rehabilitation and construction works. Moreover, improved spatial information on RPQs not only drives policy and infrastructural changes but also brings about behavioral improvements among the actors in the system (Ujuizi Laboratories, 2018). This information can help transporters make well-informed route plans to reduce the overall cost of transport. Traders benefit from the realization of their profit and loss as well as streamlining their logistics for efficiency. An emergent result of this information would be lowering PHLs, by a part attributed to the transport stage of the supply chain. Prospectively, objective sources of information on road conditions that can be obtained efficiently and conveniently pave the way for further research and development. For instance, it can be used in developing PHL models that can make of RPQ information to estimate losses with multivariate analysis, as proposed by Ujuizi Laboratories (2018), or logistical models to evaluate transporting costs of the road corridor. By extension, such works integrate well with the effort of Nelson et al. (Nelson et al., 2006) in developing a high-quality publicly available global road information database.

3.3. Organization of the Thesis

This thesis document is organized as follows:

- Chapter 1 Introduction: This chapter consists of the background, motivation, and conceptual framework of this study.
- Chapter 2 Literature review: This chapter includes the various research works, state-of-the-art, and the existing methodological gap related to assessing and mapping road pavement quality.
- Chapter 3 Research objectives and questions: This chapter includes the problem statement and outlines the guiding objectives and related research questions.
- Chapter 4 Methodology: This chapter presents the methodology undertaken to achieve the objectives of this study.
- Chapter 5 Results and discussion: The findings of the thesis are presented in this chapter, along with an analysis of the findings.
- Chapter 6 Conclusion and recommendations: This chapter presents the insights generated from the study and the suggestion for future related works.
- Appendix: This includes the supplementary materials that support the thesis document.
- References: This presents the bibliography of the works cited within the document.

2. LITERATURE REVIEW

This chapter reviews methodological approaches used by research works in assessing and monitoring road conditions. The approaches were categorized into standard (traditional), crowdsensing, remote sensing, and machine learning-based to illustrate the advancement in the field, the state-of-art, and the challenges faced in those approaches. In the end, a conclusive remark will be drawn, describing the existing technological gap in road condition assessment, which serves as the logical basis for this research.

2.1. Standard road condition monitoring methods

Standard road condition information, such as roughness and surface distress, is typically collected using specialized vehicles fitted out with high precision laser and position sensors (Laubis et al., 2019). This method is the automated/semi-automated and naturally preferred alternative to manual profile measurement, e.g., rod and level, due to its relative efficiency and lower labor costs (Wang, 2018). These approaches provide absolute profile measurement at high accuracy representing IQL I or II (Class I or II). Nowadays, response-type road roughness measurement systems (RTRRMS) such as those using accelerometers and transducers have gained interest in indirectly offering road roughness indicator information from vehicle response measurements. Since these response-type methods do not measure absolute profiles directly (only correlational) and often have lower accuracy than the earlier standard methods, the information obtained typically belongs to IQL III (sometimes arguably IQL II). Quantitative measurements obtained through standard methods are often aided with visual in-situ surveys based on human observations in small-scale inspections (Fagrhi & Ozden, 2015). Surface distresses, in particular, are assessed through video distress analysis, transverse profilers, and commonly with on-site visual surveys (Bennett et al., 2006).

Although these techniques offer accurate assessment, they are expensive, time-consuming, and inefficient, especially in developing countries and extensive road assessments (Cadamuro et al., 2018; Fagrhi & Ozden, 2015). As a result, these inspections are made at long intervals or not at all. This coarse temporal granularity of the information limits its utility for determining efficient road maintenance strategies by road authorities and real-time dissemination to road users (Laubis et al., 2019). Alternative systems equipped on vehicles and specialized for detecting surface defects (distresses), such as ground-penetrating radar, photo, or video cameras, thermal or acoustics, Light Detection and Ranging (LiDAR), and Terrestrial laser scanning (TLS), have also been used. However, none satisfies the technical (accuracy and practicality), cost efficiency, or information frequency criteria (Schnebele et al., 2015).

2.2. Crowdsensing approaches

The widespread usage of smart devices and modern vehicles equipped with numerous sensors enabled several crowdsensing-based approaches (i.e., users as data sources) to complement or replace traditional road monitoring approaches (Eriksson et al., 2008; Laubis et al., 2019). Crowdsensing methods have crossed into practical use cases (Lars Forslöf & Hans Jones, 2015; Ujuizi Laboratories, 2018), leveraging their novelty in using the frequent travels of road users to allow extensive coverage. There has been much research into using smartphones for detecting specific road irregularities like potholes, speed bumps, sunken manhole covers, and so on (Eriksson et al., 2008; Mohan et al., 2008; Rajamohan et al., 2015). Wang (2018) and Laubis et al. (2017; 2019) provided an extensive summary of the various smartphone-based application and crowdsourcing efforts in collecting road quality information. Typically, these approaches detect road defects through the various data channels of smartphones (e.g., GPS, accelerometer, digital camera, etc.) or determine road roughness indicators such as IRI through correlations. They aim to offer road condition information at IQL III or II. With the recent advent of machine learning algorithms, the approach shifted to regression (determining precise road roughness values at IQL III or II) and classification (road quality classes information suitable for IQL IV) solutions common in machine learning methods. The use of machine learning facilitated the self-calibration of new vehicles into a crowdsensing system (Laubis, 2017). This aspect made the crowdsensing technology more reliable, robust, and practical in road condition monitoring with high spatiotemporal coverage. Moreover, it enabled a more effortless fusion of multi-sensor data accessible through smart devices.

These approaches, however, are often challenged in achieving reliable information, especially when concerned with high-level decisions such as the rehabilitation of road networks (2019). Variations in sensor and GPS accuracy (which often relies on cellular network systems) across devices pose questions in the reliability of crowd sensed road conditions information (Cadamuro et al., 2019). CHEETAH[1], a smartphone application developed by Ujuizi Laboratories (Ujuizi Laboratories, 2018) for PHL and road pavement quality monitoring, tried to improve the reliability of an existing road anomaly detection algorithm through user validation of identified road defects facilitated by artificial intelligence (AI). Nevertheless, the challenge of system adoption and crowd motivation would persist to the detriment of the coverage and robustness of the method. Moreover, the difficulty of calibrating for new perturbations, which can be infinite depending on driver behavior and other conditions, has not been fully addressed and can cause wrong unrelated measures.

2.3. Remote Sensing Approaches

Typically, remote sensing, in this context, would refer to any surveying method that does not require physical contact with the road surface, and this would include vehicle-mounted approaches that do not require contact (Schnebele et al., 2015). However, the term in this study excludes vehicle-mounted methods and accordingly focuses on unmanned aerial vehicles (UAVs), airplanes, and satellites as remote sensing platforms. Remote sensing techniques use the wide range (spectrum) of electromagnetic radiation to gather information in various ways depending on the region of spectrum used. Out of the various remote sensing data collection methods, optical imagery (Multispectral and Hyperspectral imagery) and Synthetic Aperture Radar (SAR) have been commonly used in road condition assessment. Multispectral and hyperspectral imagery utilize the visible and infrared range of the electromagnetic spectrum to collect images with multiple channels (bands) representing different spectrum regions. The difference between the two is that the latter provides denser information with many bands at smaller spectrum intervals. SAR, on the other hand, uses the microwave range of the spectrum.

---

[1] CHEETAH is an acronym for Chains of Horticultural Intelligence; towards Efficiency and Equity in Agro-Food Trade along the trans-Africa Highway.

Despite the importance of ground-based methods for road quality monitoring, remote sensing approaches have also emerged as suitable supplements and alternatives for this task. Through remote sensing, it is possible to collect ground information over broad spatiotemporal coverage rapidly. With the recently increasing availability and accessibility of open sources and higher quality commercial remote sensing products, the utility of remote sensing in monitoring conditions of infrastructures has significantly improved. Early works in road condition monitoring using remote sensing relied on developing a relationship between spectral signatures obtained from hyperspectral imagery and road pavement quality identifiers. Herold & Roberts (2004) pioneered the development of this relation. They found that pavement aging and erosion of asphalt mix causes a general increase in reflectance (albedo) and changes in small-scale absorption features. Accordingly, hyperspectral reflectance features have the potential to be used as a road condition indicator. Following these findings, several investigations (Abdellatif et al., 2019; Andreou et al., 2011; Bridgelall et al., 2015; Mettas et al., 2015) used spectral characteristics to assess road condition and damages. In addition to their use of very high-resolution HSI imagery, these works are dependent on a collection of spectral libraries (information), which makes the methods inapplicable at scale in providing road quality information.

The advantage of HSI in offering wide spectral coverage at fine spectral resolution enabled the formulation of these relationships. However, with a higher spectral resolution, there is often a limitation in spatial resolution. Coupled with high spatial variability of viewed objects, mixed pixels, where one image pixel can represent multiple surface features (objects), are common issues in remote sensing imagery, even in high-resolution images (Small, 2003). Particularly for roads, narrow features surrounded by other land cover types, determining road conditions from image analysis poses high uncertainty due to asphalt pavement usually belonging to a mixed pixel (Pan et al., 2017). The usual approach in addressing this issue is either to use higher spatial resolution HSI obtained at high commercial cost or to increase spatial resolution at the expense of spectral information (Karimzadeh & Matsuoka, 2021; Pan et al., 2018). At lower (spatial and/or spectral) resolutions, e.g., multispectral imagery from Sentinel-2, various image processing techniques such as thresholding, morphological algorithms, and Fourier transformation, which can isolate defects from the background and enable binary interpretation, have been used to detect flexible pavement distresses (Chambon & Moliard, 2011; Schnebele et al., 2015; Singh & Garg, 2013). However, considering sharp variations of lighting and road surface in remote sensing imagery, the automatic detection of pavement distress using image processing becomes a complex and challenging task (Pan et al., 2018).

Synthetic Aperture Radar (SAR) data can also be a suitable alternative or supplement to optical imagery data for its sensitivity to surface roughness (Workman et al., 2016) and independence from atmospheric conditions like clouds, rain, snow, fog, daylight, etc. (Ager, 2013). Moreover, SAR data is not affected by mixed pixels issues, unlike optical imagery. Meyer et al. (Meyer et al., 2020) used high-resolution SAR data acquired at X-band to develop a model to classify road segments into “good road quality” and “road in need of repair” and reported an overall accuracy of 92.6%. This work realized the applicability of SAR data in assessing the quality of secondary roads. However, the implementation depended on high-resolution commercial SAR data and was challenging to replicate in other areas and conditions. Suanpaga & Yoshikazu (2010) developed a multinomial and binary logit model to evaluate highway riding quality from Phase Array type L-band SAR (PALSAR) with a resolution of 12.5 meters and achieved an overall accuracy of 61% and 87% for respective models. A common trait observed in both mentioned methods, and arguably in other works involving feature classification, is that the accuracy is bound to improve with fewer classes. This aspect, however, comes at the expense of losing information quality. Although SAR offers immense potential in extracting road quality information, because of its side-looking nature and intrinsic constraints, such as foreshortening, the interpretation of the data is more complicated than that of optical data (Karimzadeh & Matsuoka, 2021). As a result, its potential in this task is less explored, and models developed based on the data are often complicated to interpret and replicate.

2.3.1. Spectral Unmixing

Spectral unmixing is another solution to the trade-off in the spatial and spectral resolution of remote sensing imagery. A typical issue in low/medium resolution satellite images, and even in high resolution depending on what task is of concern, is that multiple ground objects can occupy a single pixel, resulting in what is commonly known as a mixed pixel. Spectral unmixing is a procedure developed to address this issue by decomposing mixed pixels into a compilation of constituent spectra, i.e., endmembers, and a set of corresponding fractions, i.e., abundance (Keshava & Mustard, 2002). An endmember represents the ground objects, and the abundance describes its spectral (visual) proportion or dominance in the pixel. There are two types of models to undertake spectral unmixing: linear and non-linear models. Their distinction lies in the assumption of how solar incident radiation reflects from the surface. The linear model assumes a one-to-one interaction between arriving photons (incident radiation) and components on a surface consisting of spatially distinct components, also described as checkerboard mixture (Keshava & Mustard, 2002). Under this assumption, if the area under view is split proportionally according to the fractional abundances of the endmembers, then the reflected radiation will convey the same proportions of the corresponding endmember (component) (Keshava & Mustard, 2002). This condition then elicits the formulation of a linear relationship between the fractional abundance of the components comprising the area in view and the spectra in the reflected radiation. Accordingly, a mixed pixel can be expressed as a linear combination of endmembers weighted by their corresponding abundances, as shown in equations (2-1) and (2-2).

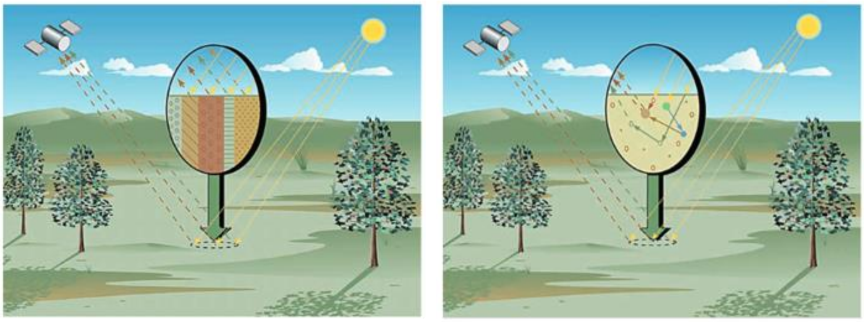


Figure 2.1 Interpretation of linear mixing (left) and non-linear mixing (right) models

Non-linear models, on the other hand, model the surface as an intricate mixture that causes multiple bounces, a condition that becomes more apparent when the size of the mixed element is small (Borsoi et al., 2020). Comparatively, linear mixing models are more widely used due to their relative simplicity, high efficiency, and transparent scientific and physical basis. At coarser resolutions and more complex ground conditions, however, the assumptions of linear models fail to encourage the use of non-linear methods.

$$x = \sum_{i=1}^{M} a_i s_i + w \tag{2-1}$$

$$X = Sa + w \tag{2-2}$$

where, x is the B x 1 spectral vector of a pixel,
a_i is the fractional abundance for endmember $i = \{1, 2, \ldots, M\}$,
s_i is the B x 1 spectral vector of endmember $i = \{1, 2, \ldots, M\}$,
w is the B x 1 observation noise vector, and
M and B represent the number of endmembers and number of spectral bands, respectively

Equation (2-2) is the matrix version applicable for images of size N pixels

In more straightforward interpretations of linear models, the abundances are usually constrained to be nonnegative and to sum to one, which restricts the spectral data to lie inside a simplex (i.e., a triangular geometry generalization) spanned by the endmembers. This interpretation provides a clear geometric understanding of the problem, as shown in Figure 2.2 for a case of 3 endmember simplex. It lays the foundation for some simple geometric-based linear unmixing methods (Winter, 1999), which rely on optimizing the fit of the spectral data within the identified endmembers simplex. However, this interpretation is often only possible under the assumption of pure pixels (i.e., pixels containing only one endmember) existing within the image corresponding to the vertices of the simplex.

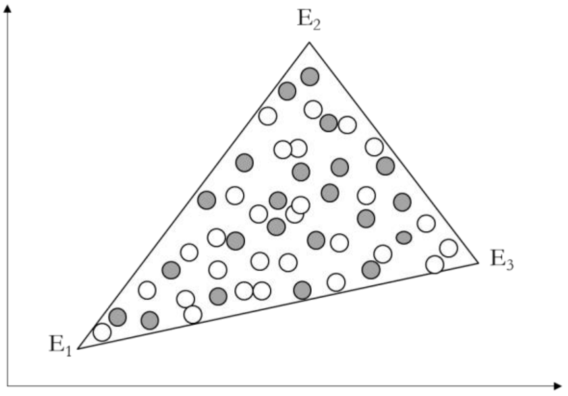


Figure 2.2 Three-endmember simplex in subspace

There has been an extensive amount of research in the field of linear spectral unmixing, all with various methods and strategies. The comprehensive review work on linear spectral unmixing methods by Borsoi et al. (2020) can be referred to for more detail on these methods, their advantages, and disadvantages. More relevantly to the topic at hand, the use of spectral unmixing methods in enhancing the potential of using medium resolution satellite imagery for road condition assessment has rarely been explored. A notable work by Pan et al. (2017) used the multiple endmember spectral mixture analysis (MESMA) method to map classes of road based on aging and distress from Worldview-2 imagery (1.84 meters spatial resolution resampled to 2 meters) with an overall accuracy of 81.71 % and Kappa coefficient of 0.77.

Remote sensing data, especially optical imagery, holds huge potential in providing road condition information at various IQLs depending on their spatial (and spectral) resolutions. Most research works have explored the use of high-quality satellite data (very high spatial and/or spectral resolution) in mapping road pavement conditions, often aiming to obtain higher IQLs information. The use of high-quality commercial images that incur a high cost for frequent monitoring deters the adoption of these techniques in resource strained regions. With the leverage of more advanced techniques such as spectral unmixing, lower resolution yet accessible imagery can offer great potential in filling this gap. Nevertheless, extraction and dissemination of this information at a convenient IQL, i.e., IQL-4, with the use of more accessible yet lower in resolution sources of optical satellite imagery, lacks attention. One notable work, in this regard, is that of, Karimzadeh et al. (2021), which used Sentinel-2 images and in-situ collected road quality data to develop a discriminant model that was able to classify road quality in Azerbaijan at an accuracy of 65% (and kappa=0.59). Although promising, the scalability of the method is constrained due to the limited transferability of the function developed for the study area, as noted by the authors.

2.4. Machine learning-based approaches using remote sensing imagery

Theoretically based deterministic models, statistical approaches, and image processing techniques have been significant in exploring the potential of remote sensing and crowdsensing data for tasks such as road extraction and road quality assessment. However, their scalability, automation, and robustness across various conditions are challenging. Appropriately, this is where machine learning (ML) methods shine. Lary et al. (2016) described ML algorithms as "universal approximators," i.e., they are able to learn underlying patterns in diverse systems from a set of training data without the need for prior knowledge about the relationship between the data and the pattern. Machine learning methods can be categorized into supervised and unsupervised. Unsupervised ML algorithms, typically consisting of clustering algorithms, can learn patterns from the input data without the need for a target output to learn from. These techniques are often used in an explorative manner to discover patterns from remote sensing observations. Clustering methods such as the K-means algorithm are widely applied in object-based image segmentation techniques. They have also been applied in spectral unmixing works to cluster locally unmixed endmembers (Borsoi et al., 2020). The use of unsupervised ML methods in remote sensing presents challenges in the interpretation of the results.

Supervised ML methods are algorithms that learn from input-output pairs to later predict the unknown desired output. They have been extensively used in various remote sensing data analysis tasks such as classification, segmentation, regression, object detection, and change detection. Supervised MLs used in remote sensing can be classified into shallow and deep ML methods (except some deep learning methods such as autoencoders, Bayesian networks, and generative models that belong to unsupervised categories). The following subsections briefly describe shallow and deep ML methods and their application in remote sensing imagery processing, particularly in road condition assessment tasks.

2.4.1. Shallow machine learning-based approaches

Before the introduction of deep learning (DL), "shallow" machine learning (ML) methods such as support vector machine (SVM), artificial neural network (ANN), and ensemble classifiers such as random forest (RF) took the focus in remote sensing data processing tasks such as image classification and change detection. The capacity of SVM to handle high dimensionality data and good performance with small training samples and the ease of use and usually high accuracy of RF models proved crucial in remote sensing-related works.

These shallow ML methods have also been used in automatic pavement defect detection in the form of an image semantic segmentation problem (Özdemir et al., 2020; Pan et al., 2017, 2018). (Ozdemir et al., 2020). Due to their ensemble nature, RF models were mostly found to perform better in these works than other shallow ML algorithms. These implementations were commonly applied on high-resolution imagery such as a multispectral image captured by unmanned aerial vehicle (UAV) and required feature selection to ensure fair segmentation accuracy. Furthermore, Yifan et al. (2018) investigated the effect of reducing image resolution on the performance of ML models used to detect potholes and cracks from UAV multispectral imagery. The classification accuracy of the models started to show a significant decline in classification accuracy over 3 cm resolution, close to the mean width of cracks, which indicated a probable threshold for a very detailed level of road condition assessment using shallow ML algorithms.

2.4.2. Deep learning-based approaches

In more recent years, with renewed interest in neural networks, deep learning (DL) methods have gained more attention in undertaking complex tasks in remote sensing such as land use and land cover (LULC) classification, segmentation, and object detection. DL models or networks are built from many layers of neurons that learn, via progressively higher-level features, to transform input data such as images into the desired output, e.g., classes (Bishop, 2006). The designation "deep" refers to a neural network that contains multiple "hidden" layers. Recently, deep neural networks became preferable because they were able to reduce the need for manual feature engineering required prior to training shallow ML models and, more importantly, because they continue to improve performance with more data.

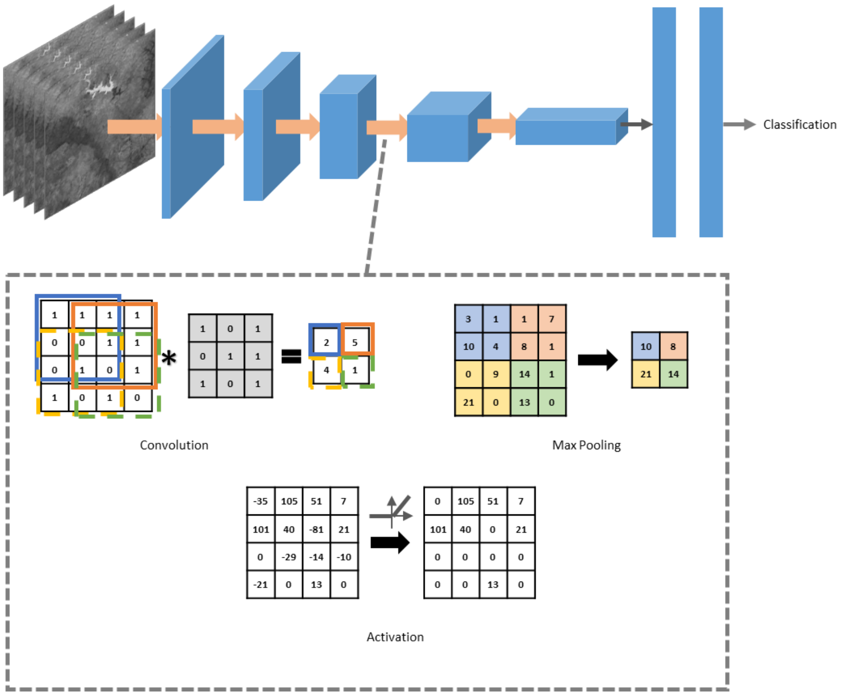


Figure 2.3 Illustration of CNN architecture adapted from O'Shea & Nash (2015)

A significant development in DL research was perhaps the introduction of convolution neural network (CNN) by (Lecun et al., 1998), which became popular after their use in the development of AlexNet (Krizhevsky et al., 2017). CNNs advanced the traditional neural network (NN) architecture by introducing hierarchical structures that included convolutional layers, pooling layers, and fully connected (i.e., regular NN layer), shown in Figure 2.3. Convolutional layers enable efficient feature learning from large unstructured data such as images through kernel-based weight sharing (Lecun et al., 1998). Through this technique, these layers generate feature maps with progressively increasing features such as edges and texture from the image, which are subjected to an elementwise non-linear transform function, i.e., activation function, to introduce non-linearity to the model. The fully connected layer (FC), where weights are no longer shared, then flattens the features and learns the classification rules (O'Shea & Nash, 2015). Pooling layers were introduced to reduce the dimensionality of the model (e.g., when images are too large) by applying spatial pooling, such as max and average pooling, over the feature maps, thereby reducing the number of parameters (Ma et al., 2019). A max pool, for instance, picks the largest element from within the applied area of the feature map (O'Shea & Nash, 2015), as shown in Figure 2.3.

More relevant to the task at hand, fully convolutional networks (FCNs) (Long et al., 2015) paved the way for the broad adoption of DL methods in semantic segmentation of images, i.e., pixel-level segmentation. While typical CNNs were built with contracting (encoder) blocks of convolution, pooling, and fully connected layers, FCN incorporates an expanding path (decoder) consisting of simple up-sampling, transposed convolution, or un-pooling layers (O'Shea & Nash, 2015). This way, they make predictions for each pixel. These networks initially attracted much attention in the medical sector and research related to road extraction from remote sensing images, which contributed significantly to the growth of the state-of-the-art in semantic segmentation using FCNs. Several researchers have used FCNs for road extraction from remote sensing images (Cheng et al., 2017; L. Gao et al., 2019; Mnih & Hinton, 2010; Oehmcke et al., 2016; Panboonyuen et al., 2017; Xu et al., 2018; Zhang et al., 2017).

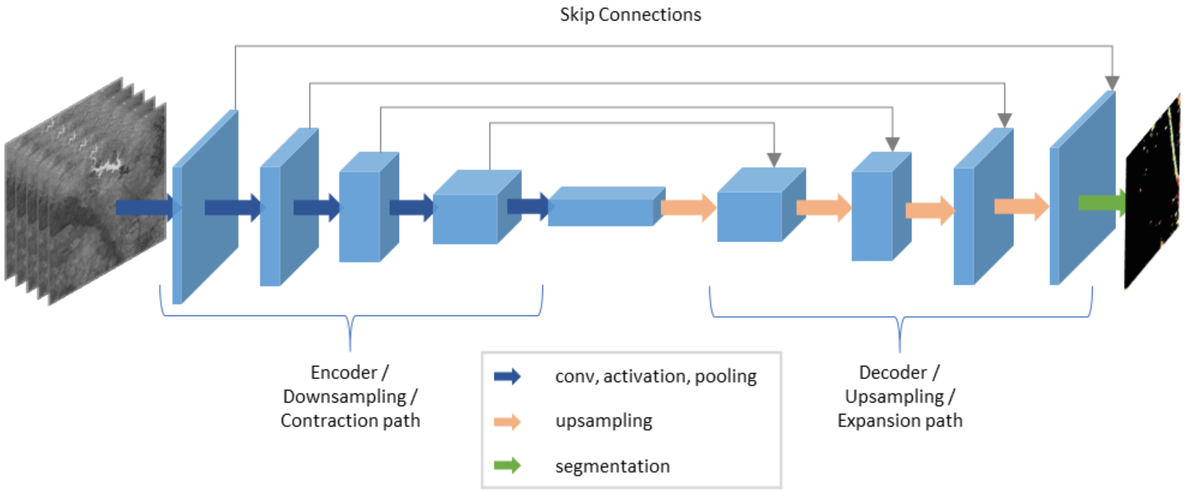


Figure 2.4 FCN architecture adapted from Long et al. (2015)

With such works, deeper and more novel models were constantly being introduced in the field of segmentation. One mentionable introduction was U-Net (Ronneberger et al., 2015), developed initially for medical research applications. U-Net addressed a persistent problem in segmentation using FCN—the loss of information during down-sampling—by including a "skip connection" that recovers spatial information by concatenating features "skipped" from each level on the encoder to the corresponding level in the decoder. Moreover, the implementation of U-Net identified the value of image augmentation in effectively training models with small training samples (Ronneberger et al., 2015). Nowadays, U-Net is the go-to architecture for most segmentation tasks either in raw form or modified for performance improvement (Zhang et al., 2017).

Given the potential to leverage the advancement of deep learning to take advantage of dense information compounded in remote sensing imagery to extract road quality, a limited body of work has investigated this capacity. Cadamuro G. et al. (2018) used extensive IRI data collected for roads in Kenya and employed various pre-trained CNNs such as AlexNet, VGG, and SqueezeNet to classify road quality from very high-resolution satellite imagery (i.e., 50 cm resolution). Approaching the problem as a categorical classification task, they obtained overall accuracies of 88 % and 73 % for binary and 5-category classification. Building on this work, Cadamuro G. et al. (2018) utilized the same dataset to predict the IRI values from the satellite imagery, i.e., regression. They made use of more advanced DL models such as LSTMs, arguing that modeling the image patches as sequential (not independent of each other) can allow the use of LSTMs, which infer information from all sections of the road in predicting for a given segment (Cadamuro et al., 2019). With this approach, they achieved an R^2 of 0.79, although, for their generalization strategy, i.e., testing for held-out data, this fell to 0.35.

At a more granular level, Oshri et al. (2018) used Sentinel 1 and Landsat 8 imagery and CNNs to predict infrastructure quality, which includes road quality, among others, based on survey collected data as a reference. In contrast to the granularity of the previous work, very high-resolution (VHR) images taken from UAV were used by Jiang L. et al. (2020) to segment pavement cracks at high precision employing a U-Net model. More recently, Brewer E. et al. (2021) were able to apply a transfer learning strategy by training various CNN architectures on comprehensive VHR imagery data for roads in the United States, for which they collected categorical road quality reference data via a phone-based platform. They applied their best model, with fine-tuning, on imagery data of Nigeria to achieve an accuracy of 94.0% in predicting road quality.

The superiority of deep learning methods in inferring more from more data offers the opportunity to leverage the increased availability of freely accessible remote sensing data such as Sentinel imagery. Although the attribute of free access often entails lower data quality (e.g., spatial resolution), deep learning methods can be designed to address these issues. For instance, by combining spectral unmixing techniques and deep learning, Alam et al. (2017) used a CNN architecture to learn features from abundance maps of unmixing and classify LULC features from hyperspectral imagery. Oehmcke et al. (2019) used deep learning frameworks to extract hardly visible road segments from Sentinel-2 time-series imagery producing good performance (35.70% F1-score for small roads) and realizing the capability to extract fine details at medium spatial resolutions. Such novel approaches can expand the limits of low/medium resolution satellite imagery to applications that can benefit from the improved accessibility of the resulting information. Regarding these constraints, in a study to assess Sentinel-2 imagery's limitations for accurate feature detection, Radoux et al. (2016) indicated that the limit for road detection is at 3m. Despite these potentials, deep learning approaches have rarely been used to extract road features from low-resolution images such as Sentinel-2 imagery.

2.4.3. Performance Evaluation

Due to the various measures of road condition assessment applied in research, the performance or accuracy of the methods used is often problematic. IRI-based assessments used in works such as that of Cadamuro et al. (2019) are evaluated based on R^2 values, typical for regression problems. Others have transformed IRI values of the reference data into categorical classifications, i.e., good, fair, and bad, based on various IRI grading systems (Brewer et al., 2021; Cadamuro et al., 2018; Karimzadeh & Matsuoka, 2021). With the possibility of obtaining this information from satellite imagery, the challenge of assessing road conditions can be posed as an image classification or (semantic) segmentation task typically assessed with the metrics of overall accuracy.

The task of extracting road quality from satellite imagery is inherently based on class-imbalanced data, whereby the target class, i.e., bad or poor road pixels/segments, is dominated by the road or non-road area of the image. As a result, standard accuracy metrics, such as overall accuracy, used for classification are rendered paradoxical, where they report high performance while the minor classes are not being predicted at all (He & Garcia, 2009). Although to a lesser extent, this effect is also apparent in road extraction from satellite imagery. Common metrics used in these works include precision, recall, f1 score, and IoU (Jaccard index), all preferred to the basic metric, overall accuracy for their intuitive behavior in class imbalance. These metrics can all be calculated from the confusion matrix, a mapping of the true labels with the predicted ones, shown in Table 2.1. The interpretation and calculation formula of these metrics is presented in Table 2.2.

Table 2.1 Binary confusion matrix

		Predicted		
	Classes	**Positive**	**Negative**	Total
True	**Positive**	True Positive (TP)	True Negative (TN)	TP + TN
	Negative	False Negative (FP)	False Negative (FN)	FP + FN
	Total	TP + FP	TN + FN	

Recently, a more interactive classification metric that plots the relationship of precision and recall values at a range of prediction probability thresholds called the precision-recall curve (plot) is gaining attention in assessing the performance of ML models. It has been proven to illustrate the susceptibility of classifiers to data imbalance and allow for a practical interpretation of model performance (Saito & Rehmsmeier, 2015). It enables model selection based on the proposed application of the results through the interpretation of the trade-off between false alarm detection (related to precision) and completeness of detection (i.e., related to recall).

Table 2.2 Description and formula of the various performance metrics used in image segmentation

Performance Metric	**Description**	**Equation**
Overall accuracy (accuracy)	▪ The percentage of instances in the dataset that is correctly predicted by the classifier ▪ misleading when data is imbalanced	$\frac{TP + TN}{TP + TN + FP + FN}$ (2-3)
Precision (Correctness or exactness)	▪ the percentage of model predicted positives that are actually true ▪ tells how much the model can be trusted when it predicts a label positive (Grandini et al., 2020) ▪ e.g., the percentage of road segments with model predicted conditions 'bad' that is actually in bad condition (Wang, 2018)	$\frac{TP}{TP + FP}$ (2-4)

Performance Metric	Description	Equation
Recall (Completeness)	▪ the percentage of true positives that are classified correctly as true ▪ it measures the ability of the model to find all the positive units in the dataset ▪ e.g., the proportion of roads with a true condition of 'good' that is also classified as 'good,' i.e., a score of 1 means that all actually 'good' road segments were classified correctly as 'good' (Wang, 2018)	$\frac{TP}{TP + FN}$ (2-5)
F1-score	▪ interpreted as the weighted (harmonic) average of precision and recall ▪ preferred to overall accuracy for its balance between precision and recall	$2 \times \left(\frac{precision \times recall}{precision + recall}\right)$ (2-6)
IoU (Jaccard's index)	▪ is a measure of similarity between labels ▪ Usually defined as intersection over union ▪ Widely used in image segmentation	$\frac{TP}{TP + FP + FN}$ (2-7)

2.5. Chapter conclusion

The following conclusions were drawn based on the reviewed works related to mapping road pavement quality.

- Traditional road quality and condition assessment methods are not efficient in providing regular and timely information on road quality.
- Higher-level IQL information on road quality obtained from granular sources can best serve in decision-making for the stakeholders considered in this study because it can be offered frequently and conveniently.
- Although crowdsensing approaches are promising, calibration difficulty, circumstantial perturbations to measurement, and ineffective system-wide adoption have limited their scalability in providing such information.
- Remote sensing data in the form of optical satellite data offer great potential in this task. However, the larger body of work on extracting road quality information from remote sensing data has focused on using high-quality information (HR and/or HSI).
- Advanced image analysis techniques such as spectral unmixing can expand the limitations of lower resolution imagery.
- Advancements in computer vision using deep learning algorithms have opened several avenues in remote sensing that have not been sufficiently explored, particularly for mapping road pavement quality using lower resolution imagery.

3. RESEARCH OBJECTIVES AND QUESTIONS

3.1. Problem Statement

Efficient, reliable, and convenient information on road pavement quality plays a significant role in reducing PHLs occurring during the transportation of agricultural products. Machine learning approaches have great potential in mapping such features from freely accessible, low-resolution images such as Sentinel-2 imagery. Up to now, very few studies have used deep learning methods to map small-scale features such as road pavement quality from Sentinel-2 imagery.

3.2. Research Objectives and Questions

The main objective of this study is to investigate the use of lower resolution optical satellite imagery and machine learning techniques in mapping road pavement quality. The following sub-objectives and related research questions are outlined to achieve the main objective.

1. To develop a machine learning model for mapping RPQ from optical satellite imagery,
    - RQ-1.1. Does a deep learning method yield better performance in mapping RPQ compared to a shallow machine learning algorithm?
2. To determine the suitable resolution for the extraction of RPQ,
    - RQ-2.1. How does using Planet imagery with higher spatial resolution affect the performance of the model compared to using Sentinel-2 imagery?
3. To investigate the added value of integrating spectral unmixing into the methodology
    - RQ-3.1. Can the performance of the models be improved through the application of spectral unmixing as a pre-processing step?

3.3. Hypotheses

The related hypotheses for the research questions outlined above are presented as follows:

RQ-1.1.

H_0 – A deep learning method performs equally or worse compared to a shallow machine learning algorithm in mapping RPQ.

H_1 – A deep learning method yields better performance in mapping RPQ compared to a shallow machine learning algorithm.

RQ-2.1.

H_0 – Using Planet imagery improves the performance of the models in extracting RPQ compared to using Sentinel-2 imagery.

H_1 – Using Planet imagery yields equal or worse performance in extracting RPQ compared to using Sentinel-2 imagery.

RQ-3.1.

H_0 – Applying spectral unmixing as a pre-processing step does not improve the performance of the models in mapping RPQ.

H_1 – Applying spectral unmixing as a pre-processing step produces improved performance in mapping RPQ.

4. STUDY AREA AND DATA

4.1. Study area

The study area for this investigation covers the road corridor from Tema (Ghana) to Ouagadougou (Burkina Faso). Otherwise known as the "central corridor," this corridor passing through Kumasi and Tamale is more commonly used and takes less travel time than the alternative shorter route via the "eastern corridor" due to the poor road quality of the latter (Roche, 2014). The road, reaching up to 1,000 km, serves as the sole surface transport route for cargo from Tema to Ouagadougou (TCBoost & WA Trade Hub, 2010). As can be seen in Figure 4.1, a large part of this corridor lies inside Ghana. This corridor is one of the eight key land corridors that link coastal city ports with major inner landlocked countries. Internal trade transactions inside Ghana hold 80-90 percent of the cargo volume transported on this corridor, while the remaining 10-20 percent is cross-border transactions between Burkina Faso and Ghana (Saana Consulting, 2016). Most of the cross-border trade involves raw agricultural produce, especially perishable products like vegetables, collected by traders from Burkina Faso and sold in Ghana among other neighboring countries (World Bank, 2019).

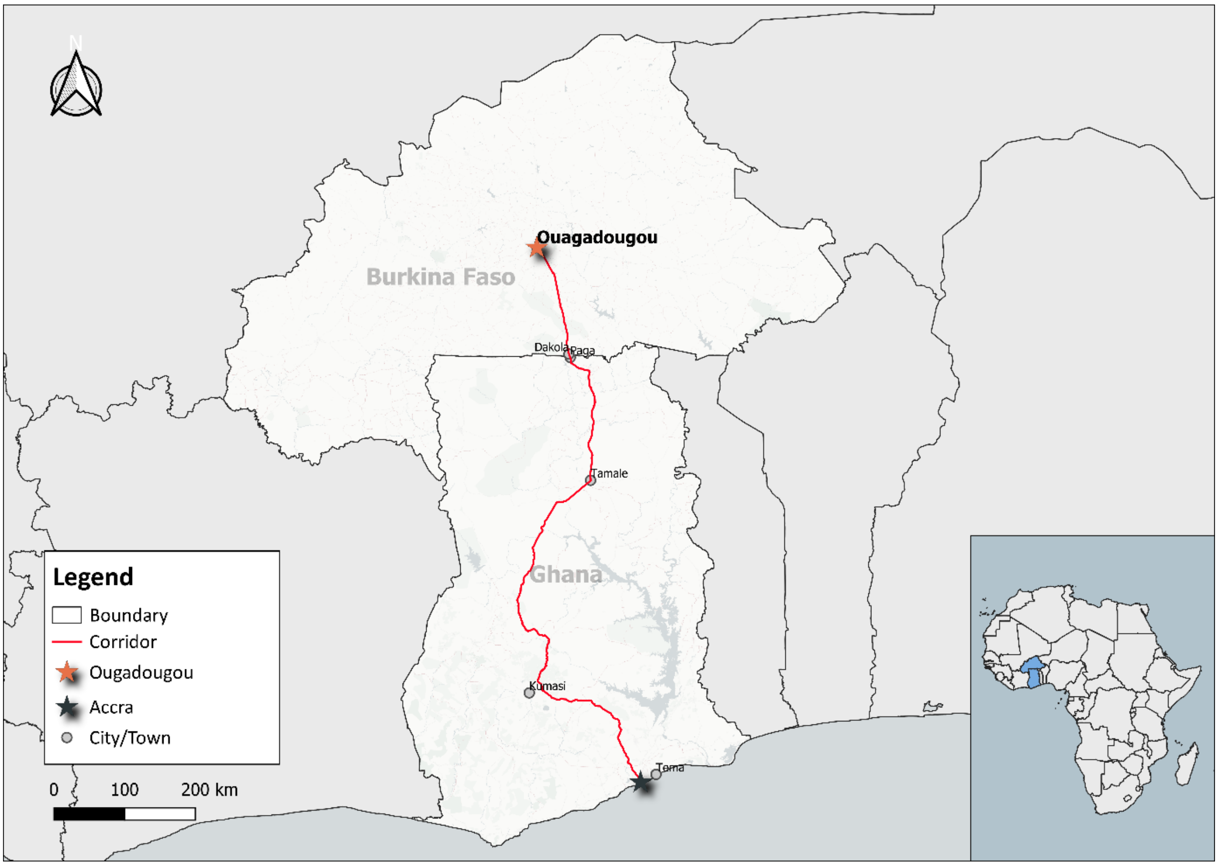


Figure 4.1 Map showing the Accra-Ouagadougou corridor

The following aspects of the corridor were adapted and summarized from an extensive study undertaken by the World Bank on regional connectivity in the area (World Bank, 2019).

- The corridor is primarily used for internal and regional trade.
- Cross-border and domestic trucking in the region remains characterized by aging trucks and own-account transport.
- The corridor section from Accra to Kumasi is the busiest route in Ghana in terms of cargo and passenger traffic; the corridor from Kumasi onward is the second busiest. Extending to Tema port, it is the sole corridor that connects Tema, Accra, Kumasi, and Tamale, which are important trade cities of Ghana.
- About 90% (over 950km) of the corridor length is in fair to good condition, while the rest (about 100km) is in poor condition. *
- It has undergone sectional improvements under various infrastructure development programs such as the Transport Sector Programs 2008-2012 and 2013-2017 (citing USAID (2017)).

* According to available data at the time of reporting

The selection of this corridor as a study area for this research was justified for two reasons. From a technical standpoint, the availability of extensive datasets regarding road pavement quality (RPQ) across the corridor collected under the CHEETAH pilot project by Ujuizi Laboratories (2018) can serve as the ground truth for this research. Secondly, from a societal problem perspective, reports from the CHEETAH pilot project indicating the prevalence of poor road conditions motivate the research attention. As per the pilot project, the road users (truck drivers), which contributed to the collection of RPQ information, reported experiencing an average of 20 to 50 road anomalies in a week. This aspect was further corroborated with the crowdsourced RPQ data (see Figure 5.2 in Section 5.1.2), indicating the prevalence of road defects across the corridor. Moreover, the continual investment in restoring and upgrading the road infrastructure—reflecting the importance of the corridor—illustrates the value of providing RPQ information regularly and conveniently for concerned authorities.

4.2. Data

4.2.1. CHEETAH RPQ Dataset

CHEETAH, an acronym for 'Chains of Horticultural Intelligence; towards Efficiency and Equity in Agro-Food Trade along the trans-African Highway,' is an app developed by Ujuizi Laboratories in an effort to provide value chain stakeholders (Transporters, Consumers, Growers, Officers from public and private agencies) a stronger voice by allowing them to express value chain flaws. The information captured by the CHEETAH app covers a wide range of fields, among which is road pavement quality (RPQ). The CHEETAH RPQ dataset refers to data collected on RPQ across the Tema-Ouagadougou corridor by Ujuizi Laboratories (2018). For the pilot project, which was carried between June 2016 and August 2017, they used the app to collect data from motion sensors and input from the users to assess road quality via crowdsensing. The motion sensor data, i.e., accelerometer, gyroscope, magnetometer, etc., was used to detect road anomalies. These detection results were then continuously enhanced by an artificial intelligence (AI) algorithm as the user encounters and confirms/adds/rejects the detected road defect (Ujuizi Laboratories, 2018). The end product data includes the detected road anomalies and the severity of the anomalies concerning road quality. RPQ in this study refers to the severity of the road defects. The data was extracted from the database via API access obtained from Ujuizi Laboratories.

The data contains GPS locations in point vectors with attributes detailing the detected road defect types, severity classes (referred to as RPQ classes in this study), and metadata related to the method of data collection and processing. The severity was categorized into three classes: 'Good,' 'Bad,' and 'Very Bad,' referring to the condition of the road segment based on the severity of the defect detected. This data, i.e., RPQ data, will be used as a reference to train the proposed machine learning models in classifying road pavement segments into the predefined classes.

4.2.2. Optical Satellite Data

Sentinel-2 optical satellite imagery data

Sentinel-2 is a multi-spectral satellite imaging mission developed under the Copernicus program. The complete mission is composed of two twin satellites: Sentinel-2A, launched in June 2015, and Sentinel-2B, launched in March 2017 (European Space Agency, 2015). These satellites fly in the same orbit but are phased at 180° to provide a high revisit rate of 5 days at the Equator (European Space Agency, 2015). Each satellite carries a multi-spectral payload (MSI) that provides measured reflected radiance in 13 spectral bands ranging from visible and near-infrared (VNIR) to shortwave infrared (SWIR) (Main-Knorn et al., 2017). The bands selected for this study and their respective spatial resolutions are described in Table 4.1. Sentinel-2 products are offered to users in two product levels: Level-2A product offering orthorectified Bottom-of-atmosphere (BOA) reflectances and Level-1C product offering orthorectified Top-of-atmosphere (TOA) reflectances. The Level-2A products can be directly used, while the Level-1C products require processing for atmospheric correction to obtain BOA reflectances. The years 2016 and 2017 were selected for acquisition to align with the reference data.

Table 4.1 Sentinel-2 image spectral bands and respective spatial resolutions

Spectral Band	**Central Wavelength (nm)**	**Bandwidth (nm)**	**Spatial Resolution (m)**
B2: Blue	490	65	10
B3: Green	560	35	10
B4: Red	665	30	10
B8: Near-IR	842	115	10
B11: SWIR1	1610	90	20
B12: SWIR2	2190	180	20

The vector of the road centerline extracted from OpenStreetMap was used as the baseline for identifying the area-of-interest (AOI) and later labeling the road section. The AOI for this study was then established by setting a buffer distance of 10 km from the road centerline. Temporally, 2016 and 2017 were selected to align with the reference data (CHEETAH data). The temporal range was crucial because straying too far (i.e., very wide range) from the time of the road condition reference data could result in a significant discrepancy in road quality apparent in the imagery to that of the reference. However, restricting the range too small meant limiting the coverage of clear and consistent imagery (i.e., cloudless with minimal spectral variation spatially) obtainable within the time range. Upon the first trial using the whole two years as the range, it was observed that in the rainy season, i.e., March/early April until mid-July and minorly September-October (Owusu & Waylen, 2013), much of the imagery in the AOI will be covered with cloud. This meant that many of the collected cloudless image tiles were gathered between January to April of 2017 except for a few (see Table A 7.2 in the Appendix). This result aligned well with the reference data and thus was used as the temporal filter for this study.

Planet optical satellite imagery data

Planet's high-resolution imagery data is accessible through Norway's International Climate & Forest Initiative (NICFI) program for free public use in scientific research and the development of policies and innovative solutions for tropical regions (Planet, 2021). This imagery data includes Planet's satellite surface reflectance mosaics that are analysis-ready. The analysis-ready specification indicates that the images are optimized for scientific and quantitative analysis by applying rigorous scene selection, atmospheric correction, cloud masking, and normalization algorithms (Pandey et al., 2020). The images consist of four bands: red, green, blue, and near-infrared, all at a spatial resolution of 4.77 m. The spatial coverage is over the global tropical region, suitable for this research since the study area lies in this region. With Level 1 access, two collections of image datasets are available:

- Bi-Annual historical archive with a temporal range of December 2015 to August 2020,
- Monthly monitoring product from September 2020 onwards.

Considering the selected timeline, the Bi-Annual data of the year 2017 was selected for this study. In particular, the July 2017 mosaic was used. The data was accessible through Planet's API using credentials obtained after signing up for the program. The mosaic was downloaded in tiles called quads that can be mosaicked together for the selected area. This data will be referred to as Planet imagery/images (or data) hereafter in this document.

5. METHODOLOGY

This chapter describes the methods that were used to achieve the research objectives established previously. The overview of the research methods workflow is shown in Figure 5.1. This research made use of two satellite imagery datasets: Sentinel-2 and Planet imagery, which require slightly different pre-processing steps (thus the different processing flows described in solid and dashed lines, respectively in Figure 5.1). Accordingly, four sets of image datasets relevant to the previously defined research question were obtained, each of which was analyzed using the two machine learning models selected for this study. The CHEETAH RPQ dataset was used to develop the label raster used for training the machine learning models. Each of the processing steps will be elaborated in the following subsections.

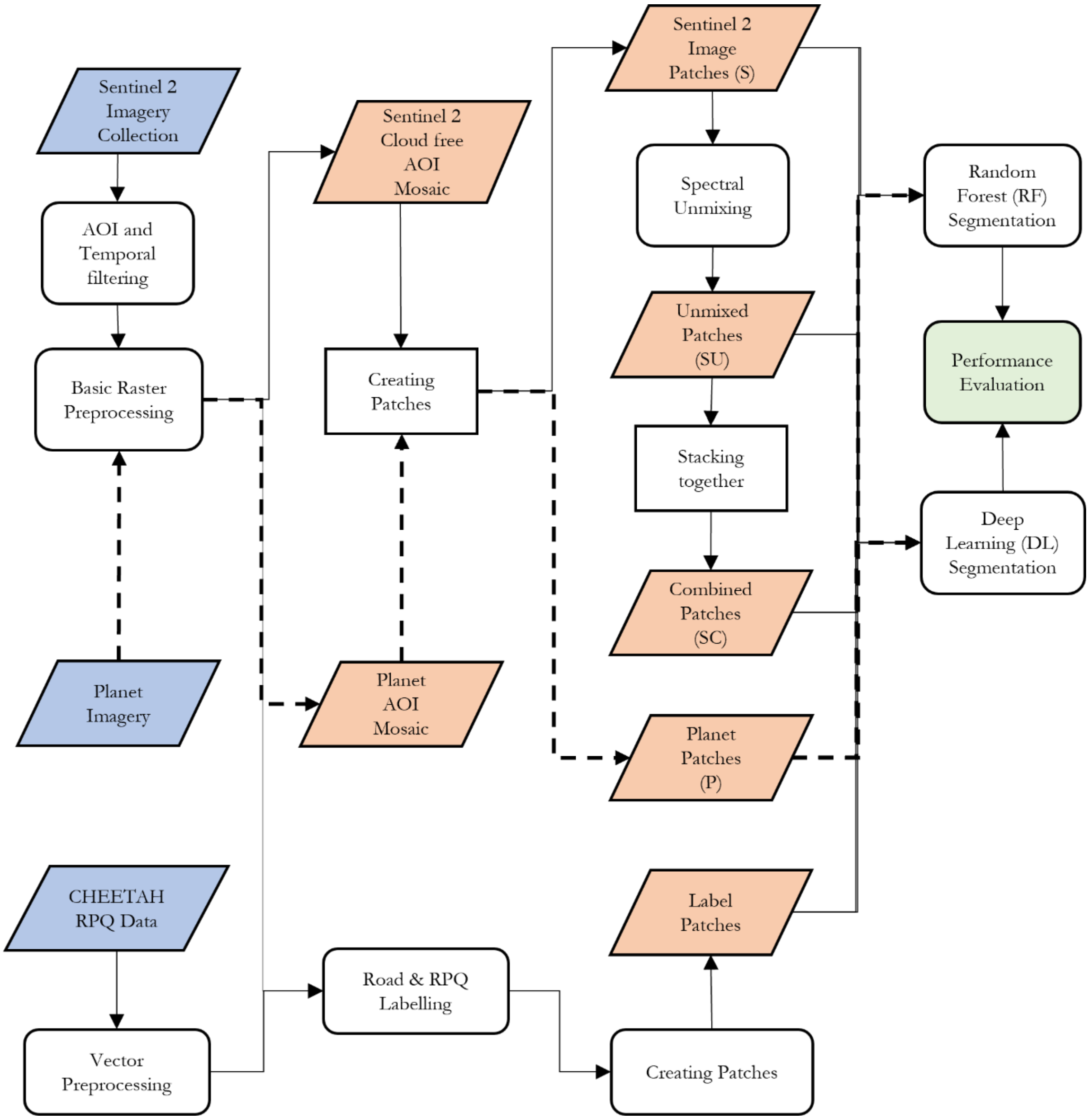


Figure 5.1 Research methodology workflow

5.1. Data pre-processing

5.1.1. Basic raster pre-processing

Basic pre-processing refers to atmospheric correction and the common geometric and raster operations implemented on the image collections to obtain a single mosaic. The Sentinel-2 images were atmospherically corrected using sen2cor (version 2.55) (Main-Knorn et al., 2017) to obtain bottom-of-atmosphere reflectance images. The configuration of the atmospheric correction as per Main-Knorn et al. (2017) is described in Table A 7.3. Digital Elevation Model (DEM) obtained from Jarvis et al. (2008) was used as additional input to estimate elevation for the correction. The method was used only for atmospheric correction. Since bands 11 and 12 of Sentinel-2 images have a lower resolution than the other bands, they were resampled to match the others using the bilinear interpolation method. The bands were then stacked together to produce multiband images. These procedures were not required for the Planet images as they are obtained analysis-ready with atmospheric correction applied to them. The images were then mosaicked and clipped to AOI. The AOI was further narrowed down to a 1 km buffer distance from the road centerline to reduce the non-road area (e.g., vegetation, bare soil, etc.) in the images, which helps improve the pixel imbalance of the road land cover class during model training. The Sentinel-2 image tiles were clipped to the AOI before mosaicking to reduce the computation load of mosaicking larger image collections. The mosaics were then normalized to center the reflectances between 0 and 1. This procedure is helpful for the performance of machine learning algorithms, particularly those based on gradient descent, because it adjusts a common scale for all the features in the data. In this case, it sets the reflectances of each band to have a standard scale of 0 to 1.

To later feed the images to the machine learning models, it was essential to split them into patches of smaller size for easier computation and practicality. The patches were created sequentially (north to south, west to east) from the mosaic image in a sliding window manner with partial overlap between each patch to enable image augmentation and increase the number of samples obtained for training the models. The choice of smaller patch size would reduce class imbalance within each patch—an issue critical in this case—and the granularity of the segmentation, at the cost of possible loss of contextual information (Hamwood et al., 2018) and increased training time. Moreover, to ensure the image size fits into the network architecture, i.e., to prevent down-sampling below 1-pixel size during feature encoding, the minimum size of patches has to be limited. Larger patch sizes were also limited due to GPU memory limitations. Accordingly, patch size of 96 by 96 pixels with 20 pixels overlaps, and patch size of 192 by 192 pixels with 40 pixels overlap were used, respectively, to generate 619 patches from the Sentinel-2 and 795 patches from the Planet mosaic images. The patch size for the Planet image was double that of Sentinel-2 since the resolution of the former is approximately half of the latter.

5.1.2. RPQ data pre-processing and road labeling

Before processing the RPQ data, road labeling was undertaken. OpenStreetMap (OSM) data was used to outline the centerline of the road. Upon visual observation on Google Earth in combination with the Sentinel-2, the vector obtained from OSM was found to have several discrepancies from the true centerline of the road (visually determined). Accordingly, these errors were adjusted. The use of the Sentinel-2 for this labeling was necessary in this case, despite its lower resolution, due to the image misalignment issues observed in the Google Earth images (Potere, 2008). A buffer distance of 3.6 meters, i.e., typical road width, was then applied to the adjusted centerline vector to define the road section. Two issues observed during labeling the road was that road width might vary in some places, and over urban areas, several roads that do not have corresponding RPQ data were within the images. These issues could detriment the quality of the reference data and the learning of the model. Therefore, some urban areas and wider road sections, which often coincided in proximity, were excluded from the imagery mosaics. These exclusions were also reasonable as the related issues were not within the scope of this study.

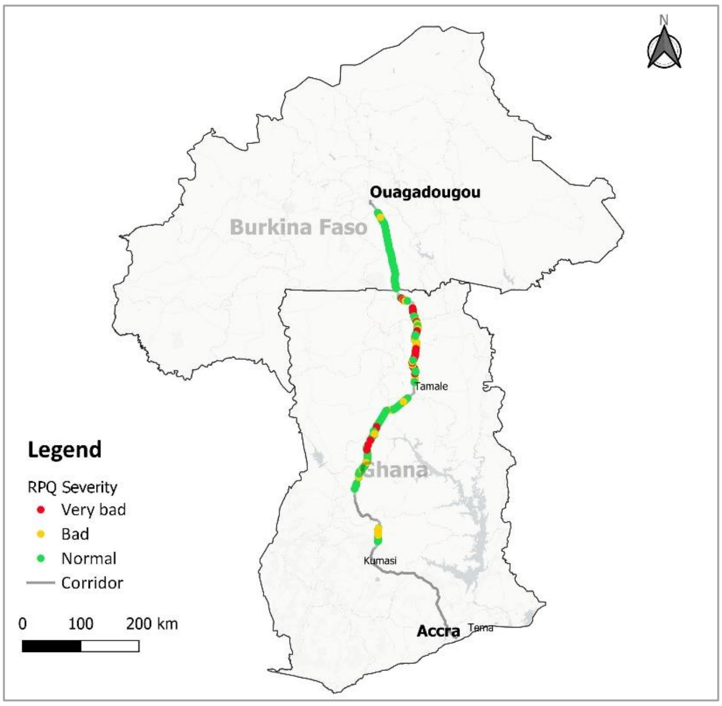


Figure 5.2 Map showing RPQ along the Accra-Ouagadougou corridor

The RPQ attribute data contained detection of the CHEETAH method, which were found impossible and/or unnecessary to detect using satellite imagery for this study. These features include labels such as police checkpoint, flooding, stops, zebra crossing, unknown, etc. Moreover, road defects type were often labeled in various descriptions. Therefore, the data was cleaned by removing unnecessary features and grouping labels. The spatial information in the RPQ data was in the form of vector points which were GPS locations of the detected road anomaly. However, these locations were often slightly inaccurate and transversely misaligned with the road. The points were snapped to the road centerline using QGIS to adjust the transverse misalignment. The configuration for this procedure includes setting the tolerance and snapping behavior. The tolerance specifies how close the points need to be to the road vector to be selected and snapped. The snapping behavior determines modification (upon snapping) of the original input vectors to match the reference vector, usually required for two or more dimensional vectors. A tolerance of 50 m was used for this case, and the points found beyond this distance were considered inaccurate and removed from the data. Since the vectors in concern here were points, the snapping behavior was set only to move endpoints with a preference of the closest point. The points were then buffered by a distance of 4 m, equal to the average width of the road, and converted to area features to account for the snapping shift and GPS inaccuracy. This conversion entails that RPQ information is expressed in sections of the road within a 4 m distance longitudinally of the point location of the detected defect. For instance, it can describe a certain 8 m (2 x 4 m) section (length) of the road as poor if a road anomaly was detected in that area of the road. However, it is important to note that this interpretation changes when these vectors are rasterized for the preparation of the label and that it also depends on the resolution of rasterization. Accordingly, for the Sentinel-2, the granularity will be at 10 m long segments of the road. Nevertheless, the choice would remain reasonable since, in satellite images, assuming road defects such as rough patches or aged road sections have a spectral influence over, at least, an area represented by a pixel(s).

5.1.3. Spectral Unmixing

The problem of mixed pixels in satellite imagery, more apparent at lower resolutions, makes distinguishing small features in the images (e.g., pothole) very difficult. Such features usually occupy a pixel together with other background features in proximity. One solution to this issue is to interpret the image at a sub-pixel level and describe a pixel as a proportional combination of the spectra of the features inside of it (i.e., endmembers). This procedure is called spectral unmixing. Image spectral unmixing aims to estimate the number of reference features (i.e., endmembers), their spectral signatures, and their proportions in each pixel of the image (i.e., abundance fractions). This study uses this technique to assess its potential in improving the extraction of road pavement quality information from satellite imagery such as that of Sentinel-2. The use of this method in various satellite image segmentation tasks, including road and road condition extraction, was discussed in this document's literature review section (see Section 2.3.1). There are numerous techniques for spectral unmixing, each with its advantages and disadvantages.

In applying spectral unmixing for this study, spectral library requirement, automation, and computational efficiency were considered important criteria, along with the method's reliability. Some unmixing methods require spectral libraries to be established, which makes the procedure contextual and less adaptable. In addition to their adaptability, the possibility of automating the procedure makes unmixing techniques that do not require spectral libraries more desirable in our case. Automating the process improves the frequency and ease with which the information produced can be disseminated to the users. Likewise, reducing computational demand is often advantageous towards reducing the cost and time of computation. However, these choices come at the cost of the accuracy of the unmixing result. The choice of unmixing method was made considering this trade-off. Accordingly, the sparsity promoting iterated constrained endmember (SPICE) unmixing algorithm developed by Zare & Gader (2007) was used to undertake the spectral unmixing procedure in this study. The SPICE method is a sparsity-promoting extension of the iterative constrained endmember extraction algorithm (ICE) (Berman et al., 2004). Due to the intricacies of the SPICE algorithm, the details of its formulation as per Zare & Gader (2007) are described in Annex 3 of the Appendix.

The SPICE unmixing method (particularly the python implementation) has four primary parameters: μ (the regularization parameter), Γ (constant controls the degree of sparsity), endmember pruning (removal) threshold, and the iteration stopping criteria. The last parameter refers to the desired change of the objective function between iterations below which the iteration stops. Optionally the iteration can be stopped by capping the maximum number of iterations. Although there is an option to insert initial endmembers (e.g., from a spectral library), the SPICE algorithm can initialize with randomly selected endmembers. Eventually, it generates endmembers—in addition to estimating their numbers—fitting the mixing model of the image without the need for a user-defined spectral library or the assumption that all the endmembers have pure pixel representation in the image (Berman et al., 2004). This capability makes the SPICE method easily automatable.

Moreover, this study uses a localized unmixing strategy by fragmenting the large mosaics into smaller regions via a sliding window approach. In this way, different sets of endmembers and corresponding abundances are extracted for each sliding window. This strategy can enhance the unmixing results by simplifying the spectral mixture problem into smaller sub-spaces with less complex mixtures and spectral variability. Spectral variability means that a feature or endmember may be represented by various spectra instead of a unique one because of changing illumination conditions, the intrinsic variability of the feature, and atmospheric effects (Borsoi et al., 2020). Via localized unmixing, there is a possibility to collect different spectra for a single feature as viewed in different windows, thereby accounting for the spectral variability of that feature. This approach is also computationally more efficient than unmixing larger-size images. However, the challenge of this strategy comes in globalizing (generalizing for the whole image) the local unmixing results. Typically, this generalization is made by grouping all the local endmembers into bundles using a clustering algorithm. The local proportional contributions (i.e., local abundances) are then summed up in each pixel if they belong to the same class (bundle). Although this approach is intuitive in addressing spectral variability, it is important to note that the results of the unmixing will significantly depend on the clustering reliability. Considering clustering is a typically unsupervised method, the resulting clusters are often difficult to interpret, thus adding uncertainty to the unmixing procedure.

Endmember clustering

Clustering is a commonly unsupervised algorithm of grouping similar objects into clusters, making it possible to discover similarities and differences between the objects and obtain the information contained in them. Clustering is a valuable technique in remote sensing data processing and has been widely used in segmentation, classification, feature selection, and change detection application of remote sensing images. More relevantly, clustering has been effectively used in unmixing satellite imagery through bundling extracted endmembers (Borsoi et al., 2020), even for road condition assessment (Pan et al., 2017). Commonly distance-based clustering algorithms such as k-means (Lloyd, 1982; MacQueen, 1967) and fuzzy c-means (Bezdek et al., 1984) are applied for image processing. However, when it comes to features that display strong spectral inseparability, in this case, poor and good asphalt road, distance-based methods often fall short in identifying separate clusters. Probabilistic mixture models that apply clustering based on the assumption that the data can be explained as a mixture of a specified number of distributions have been found to be intuitive in addressing this issue. With this assumption, it is possible that separate distributions (i.e., separate clusters) can overlap, which indicates that these approaches do not assign hard clusters. Their clustering is instead based on a probability that a data point belongs to a certain distribution, often described as soft clustering.

The mixture model with multivariate Gaussian distribution, usually called Gaussian mixture model (GMM), has been widely used in various remote sensing image processing tasks (Ait Kerroum et al., 2010; Ari & Aksoy, 2014; Çelik, 2011; Tadjudin & Landgrebe, 2000; B. Zhao et al., 2016). As the name indicates, GMM assumes a mixture of Gaussian (normal) distributions of data points. This model is useful in clustering data, with elliptically shaped groupings—an often difficult task for distance-based clustering algorithms—, which is more apparent in remote sensing images. GMM defines clusters by estimating the parameters (i.e., mean, variance/covariance, and size) of the distributions in the data by maximizing a certain likelihood function, i.e., the maximum likelihood estimation method (MLE). One of the most widely used algorithms for this estimation is the expectation-maximization (EM) algorithm (Dempster et al., 1977).

An inherent challenge in using GMM is its complexity when it is applied to cluster high dimensional data (Thiesson et al., 2001). The model tends to show confusion in identifying the true structure of clusters when the data has a large number of features, some of which do not contribute to the cluster delineation (i.e., "noisy features") (Maugis et al., 2009). Although the dimensionality in our case is relatively low, the intrinsic inter-band correlation of Sentinel-2 imagery (J. Chen et al., 2017) results in significant difficulty for clustering by adding noise. Clustering tasks have often been done after a feature selection process (or dimensionality reduction) to address this issue, whereby only important features are chosen to represent the data for clustering. Due to its simplicity and computational efficiency, principal component analysis (PCA) has been widely used as a dimensionality reduction method. Moreover, the number of features (components) to represent the data does not have to be known a priori since it can be determined intuitively based on the desirable explained variance of the data under the new components. This aspect enables automation under the desired threshold of explained variance.

The main parameter in this clustering is the number of components (clusters), which, typical of unsupervised clustering, is often a difficult task since the algorithm will always prompt more components. There are several measures developed to select the number of components (Akaike, 1972; Rissanen, 1978; Schwarz, 1978), among which the Bayesian Information Criterion (BIC) (Schwarz, 1978) is the most commonly used. BIC is a model selection criterion derived from the Bayesian probability concept and suited for models based on the maximum likelihood estimation framework. It is preferable to its alternatives because of two reasons. Firstly, it optimizes the trade-off between model performance and complexity through a strong penalization of more complex models (Bishop, 2006). Secondly, its basis in Bayesian probability entails that the probability that it will select the correct model increases as the sample data size increases (Hastie et al., 2009). However, for smaller and less representative datasets, the criterion will be more likely to choose undesirably simple models (Hastie et al., 2009). Normally, the model (or the number of components) with the lowest BIC value will be selected. However, according to Zhao et al. (2008), the BIC value can often decrease continuously such that a minimum cannot be found at a reasonable number of clusters. It is, therefore, recommended to locate the knee point of the “number of clusters vs. BIC value” curve instead (Q. Zhao et al., 2008). The knee (or elbow) point is the point at which the gradient (slope) of the curve visibly changes from high to low. In this case, it is the point after which there is no value in increasing the number of clusters and hence is the optimal number of clusters for the model. Although this point can be roughly identified through visual identification, it is preferable to use a theoretical basis for consistency and automation. Consequently, the method developed by Satopää et al. (2011) was used to find the knee point, preferred to other approaches for its versatility and ability to fit on partial data (i.e., “online algorithm”).

Based on the choices mentioned above, the following strategy was utilized for the unmixing procedure in this study. An important note here is that the unmixing pre-processing, although initially proposed, was not applied to the Planet image. The spectral information limitation of Planet images, i.e., only four bands, would minimize the effectiveness of unmixing, e.g., limiting the number of discoverable endmembers. Accordingly, the following procedures are only applicable to the Sentinel-2 imagery data.

i. Since the mosaic image was partitioned into smaller patches for machine learning image segmentation, which suited the localized approach selected for unmixing, the local spectral unmixing was applied directly on these patches.
ii. The patches were unmixed using the SPICE method initialized with the parameters listed below obtained based on the recommendation of Zare & Gader (2007) and modified empirically. The selected parameters, their recommended values, and the remarks on why they were changed if so are described in Table A 7.4 of the Appendix.
iii. The unmixing process outputs the spectra vectors of endmembers and images with the proportion of each endmember in each pixel (abundance map) for each patch. The abundance maps have bands representing the endmembers identified in the patch.
iv. The endmember spectra of all the patches were collected together to form a large library of endmembers.

v. Dimensionality reduction was applied to the endmember spectral library using the PCA method parameterized to have 95% of the data variance explained, which resulted in three components.
vi. The endmembers were then clustered using the Gaussian mixture model optimized by the EM algorithm. BIC values were used to optimize the model and determine: the covariance type and the number of components (clusters). The BIC knee point for each covariance type model was calculated and compared to select the best model, i.e., "full" covariance with five components. The comparison of the knee points for each covariance type and the plot describing the knee location of the best covariance type are shown in Figure A 7.1 and Figure A 7.2, respectively.
vii. Accordingly, five clusters were defined for the endmember spectral library.
viii. Based on these defined clusters, each patch's corresponding abundance maps (image representing the endmember proportion in each pixel) were rearranged and/or summed up to obtain abundance maps that had consistent band-wise structure through the patches. This patch images dataset named hereafter in this document as unmixed Sentinel dataset or SU for short was established as input data to the machine learning model in relevance to the third research question of this study.
ix. These images were stacked together (band-wise) with their corresponding Sentinel-2 normalized patch images to explore further the potential of the unmixing step in the improvement of the machine learning predictions, again with regards to the second research question. This dataset will be, henceforth in this paper, referred to as the combined Sentinel dataset or SC for short.

5.2. RPQ segmentation using machine learning

With the aim of addressing the second research question of this study, a deep learning method in the form of a fully convolutional neural network (FCN) model was selected and compared with a shallow machine learning model in the task of mapping road pavement quality posed as a satellite image segmentation. The following subsections explain the shallow and deep learning models investigated in this paper.

5.2.1. Shallow machine learning: random forest classifier

Random forest (Breiman, 2001) was selected for the baseline shallow machine learning approach. Random forest (RF) is an ensemble-type learning algorithm built from a collection (forest) of a certain number of decision tree classifiers. Each tree is trained on a different bootstrapped sample of the training data, in which these samples are selected randomly with replacement to increase diversity among the trees (Breiman, 2001). It also applies a similar bootstrap approach for the features used to find the best split by selecting a random subset of the original feature set at each node of a tree. The splits are then tested based on a selected feature value test, e.g., maximization of the Gini index. The data is accordingly split recursively until all the nodes are pure (i.e., all samples in each node belong to the same corresponding class) or when a user-defined criterion such as maximum depth is met. The predicted class of a sample point is determined from every tree in the forest and assigned based on a majority vote of all the trees. These characteristics of random forest strengthen its robustness and generalization compared to learning algorithms. It also allows the evaluation of the importance of every feature as a predictor, which is insightful for model tuning.

The hyperparameters to optimize random forest classifiers include the number of trees (or estimators), split quality criterion, maximum tree depth, maximum features considered for the best split, and whether to bootstrap the training data or not. Among these parameters, the number of trees is the most important one, which significantly influences the accuracy and computation efficiency of the model. The split quality criterion refers to the test function used to measure the quality of splits. In the Sklearn framework (Buitinck et al., 2013) used here, there are two options for this parameter: the Gini impurity and information gain (entropy). Maximum depth is the maximum number of levels in a tree. Sklearn has various functions to calculate the maximum features for best split based on the number of features in the data, the selection of which was set to be a hyperparameter for tuning in this study.

Although RF models are typically trained for the whole batch, this study uses a mini-batch approach in training the RF model due to computational limitations. This approach adds more estimators (trees) to the model after each batch. It fits the new model, which retains aspects of the previously fitted model as initializers to the subsequent batch. This process is repeated until the entire data is fitted. The number of estimators to add and the mini-batch size were set to be hyperparameters for tuning.

Accordingly, the model was tuned to select the best combination of these hyperparameters mentioned above. A random search algorithm was used to sift through the hyperparameter randomly until a limited number of trials and report the hyperparameter combination that resulted in the maximum performance score. The selected hyperparameters are reported in Table A 7.6 of the Appendix. The model was then trained using the selected hyperparameters on the respective dataset.

Since random forest models cannot extract features themselves, unlike DL methods, a feature extraction process is applied prior to feeding the data for training/tuning. Basic features of intensity, edge, and texture were computed for each band of the images at different scales by applying Gaussian blurring with a range of kernel values used to average local neighborhoods. The Gaussian blurring applied at different kernel values before the computation of the features allows them to be created at different spatial scales, i.e., multiscale feature creation. This technique enables the model to potentially learn from features at various spatial levels.

5.2.2. Deep learning: U-Net

This research utilizes the U-Net architecture (Ronneberger et al., 2015), a deep learning model widely used in semantic segmentation, for road pavement quality segmentation from optical satellite imagery. U-Net is a deep encoder-decoder type FCN, consisting of a contracting or down-sampling path (encoder) to capture context and an expansive or up-sampling path (decoder) for object localization. The novelty of this architecture was in the skip connections between the encoder and decoder paths, through which feature maps from the encoder are copied, cropped (if necessary), and concatenated to the corresponding layer in the decoder.

A slightly modified architecture—from the original—was applied for this study. The encoder of this architecture, shown in Figure 5.3, consists of repeating blocks of two 3x3 convolutions, each followed by a batch normalization, a leaky rectified linear unit (LeakyReLU) activation, and ending with a 2x2 max pooling operation with a stride of 2 for down-sampling (except for the last one, i.e., the bridge, which does not apply max pooling). The feature channels (the number of feature maps) are doubled with each down-sampling block. The batch normalization and LeakyReLU (with a negative slope, i.e., α, of 0.3) were applied here, instead of the direct ReLU, to improve the training performance (Ioffe & Szegedy, 2015) and prevent the dying ReLu problem(Lu et al., 2019; Maas et al., 2013), respectively. Dying ReLu is a problem in training deep neural networks when ReLu activations only output zero for any input, effectively killing the corresponding neuron and reducing model learning capacity (Lu et al., 2019). LeakyReLU prevents by mapping the zero outputs to values at a certain slope (α). The decoder is also formed from two repeating similar blocks. These blocks consist of a 3x3 transposed convolution (inverted convolutions) with a stride of 2, such that it halves the number of feature channels, and a 3x3 convolution followed by batch normalization and LeakyReLU. After the transpose convolution, the halved feature maps are concatenated with their corresponding feature maps (cropping was not needed here as their sizes matched) from the encoder. The final layer after the decoder is a 1x1 convolution with a SoftMax activation used to map each feature vector to the number of classes for segmentation. Additionally, the input image size of the model was adjusted to fit the size of the patches, i.e., 96x96 pixels for those corresponding to Sentinel-2 and 192x192 pixels for those of Planet.

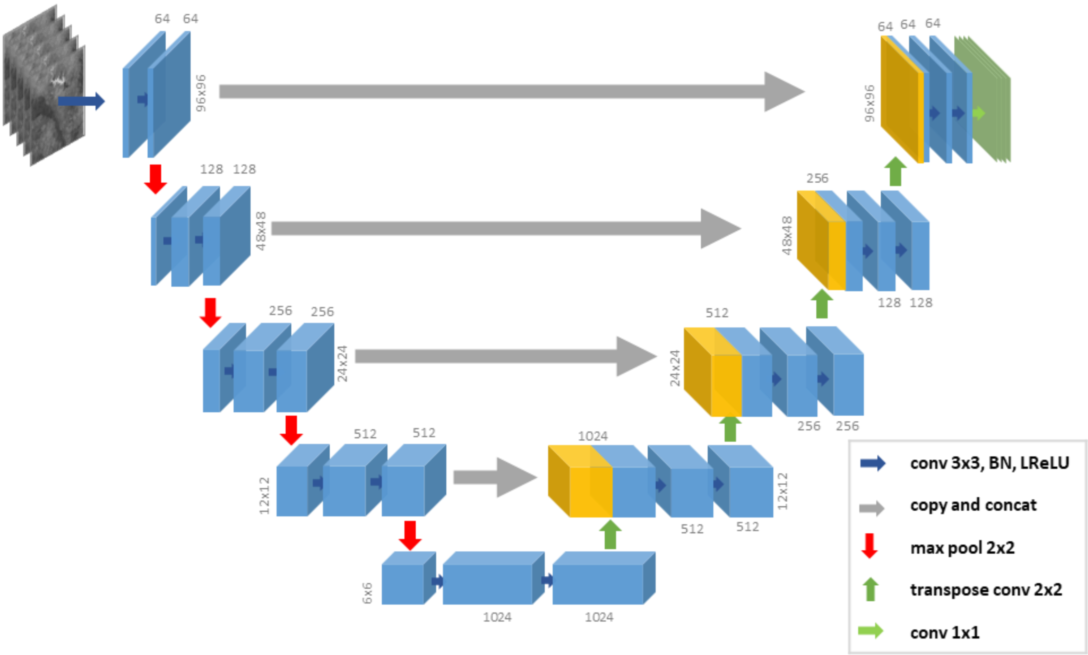


Figure 5.3 U-Net architecture used in this study adapted from Ronneberger et al. (2015)

5.2.3. Data imbalance and focal loss

The size of roads in the satellite images is too small such that their pixel-level representation in the data causes a class imbalance in the target labels. This aspect is even more apparent in road quality labels, as shown in Table 5.1. The data shows a drastic imbalance between the RPQ label classes and the background non-road class, even with under-sampling the background label by reducing the AOI and patch size. A common technique applied to address this issue in DL-based small object detection works is to modify the loss function such that it encourages the detection of ‘hard’ classifications (underrepresented class).

Table 5.1 Label class balance

Label class	**Class to total ratio**
Non-road	0.989775
Good road	0.009648
Bad road	0.000394
Very bad road	0.000184

Focal loss (FL), a loss function proposed by (Lin et al., 2017), has been found to be effective in achieving good performance in class imbalanced small object detection tasks. It is a modification of the widely used Cross-Entropy loss (CE). When training a model for data with a large class imbalance, the loss gets overwhelmed (dominated) by the easily classified majority class (background), which determines what the model learns to predict by influencing the gradient (update of the weights). FL modifies the CE function by including a class weighting factor (α_t) and, more importantly, a modulating factor. The class weighting factor improves prediction on imbalanced data by weighting the loss based on the predicted class. The modulating factor improves training on imbalanced data by encouraging hard classified examples (predictions with low probability, e.g., bad road class) and the opposite for easily classified examples (e.g., background object such as a non-road class in this case) (Lin et al., 2017). These two factors were considered hyperparameters, and their values were set based on model tuning.

5.2.4. DL model training

To address the limitation of available data, four sets of image augmentations were selected and applied randomly from six types of augmentations: 90^0, 180^0, and 270^0 rotations and horizontal, vertical, and transpose flipping. These augmentations are reasonable to make since the expected road orientation can be at any angle. The model was trained using Adam optimizer (Kingma & Ba, 2015), selected for its ease of implementation and computational efficiency. An initial learning rate of 0.00001 (i.e., 10^{-5}), obtained via hyperparameter tuning, was used to train for a maximum of 100 epochs, which were capped off via early stopping on the plateau of validation accuracy with a tolerance of 15 epochs (smaller value was used for tuning). This strategy stopped the training and saved the best model weights when the model was no longer improving for 15 epochs, preventing overfitting and reducing training time. The learning rate was reduced stepwise throughout the training at a rate of 0.5 (i.e., halving the learning rate) every 20 epochs based on tuning results. A different learning rate schedule was used for the Planet images with a rate of 0.5 every ten epochs (see the section below for reasoning). Both L1 and L2 regularization techniques were applied since the model had a strong tendency to overfit. These regularizations will penalize larger weights within the model, hence reducing the complexity of the model, i.e., its tendency to overfit.

5.2.5. DL model tuning

The U-Net model was tuned to develop a model of the best possible hyperparameters with due consideration of computational limitations. Since deep learning models have several hyperparameters, the tuning was done in a divided manner. As per Goodfellow et al. (2016), DL model development should be started from the simplest implementation, ensuring the model's capability to overfit on the training data, after which improving parameters can slowly be included to get the best model. This approach ensures the expected functionality of the model. Accordingly, the model was tested and confirmed to overfit on the training dataset. Next, a suitable initial learning rate was established empirically, with default settings for other hyperparameters, balancing between the overshooting problem of high learning rates and the slowing effect of low learning rates. Then the first group of hyperparameters, i.e., L1 and L2 regularization rates and focal loss function hyperparameters, were tuned using the selected learning rate and default settings for other hyperparameters. Lastly, the model was refined by tuning the batch size and learning rate schedule (i.e., decay rate and steps). Due to processing limitations, the batch size was limited to lower numbers (i.e., at the cost of time). This strategy for tuning the model was found reasonable because the strength of influence in the model performance of each of those hyperparameters follows the same order, and thus makes tuning hierarchical and easier to understand hyperparameters' influences. The tuning of the last two groups was automated through a random search algorithm. This algorithm goes through the multiple parameter combinations within the hyperparameter space randomly until the limit of maximum trials, i.e., 7 and 12 in this case, respectively, for each group of hyperparameters. The best hyperparameters obtained after tuning are shown in Table A 7.7 of the Appendix.

5.3. Performance Evaluation

Several ML performance evaluation metrics, particularly those relevant to image segmentation tasks, were considered to comprehensively address the research questions and investigate the potential of the RPQ mapping method proposed in this study. An inherent challenge distinctive to RPQ mapping (semantically segmenting) is class imbalance within the target data. This imbalance is caused by the minuteness of the road defect features (in pixels) within satellite images. Evaluating models under heavy data imbalance is a paradoxical undertaking. The major class(es) can overshadow the minor ones and display near-to-perfect performance in several metrics, such as overall accuracy. Even though directly using the confusion matrix to assess the misclassifications is insightful, it raises difficulty in comparing results with other works. Based on the literature review, precision, recall, and F1-score are known to be particularly useful in evaluating the performance of machine learning models on imbalanced datasets. Jaccard index or intersection-over-union has also been used widely specific to segmentation evaluation.

Balanced accuracy weighted by per-class rarity weights was used during training to assess performance over the validation set. Balanced accuracy is the class averaged (also called macro averaged) recall. When weighted with the rarity (inverted frequency), this accuracy score rewards detection of the rare class more than others (Brodersen et al., 2010; Cohen et al., 2006; Gupta et al., 2020). The use of this metric helped optimize the model to be sensitive in detecting bad and very bad road pixels.

During testing (evaluation), class-wise metrics and scalar summaries were used to assess the generalization of the models. The scalar summaries were used to report and compare experiment accuracies and select the best performing model. The metrics used for this report are precision, recall, F1-score, and IoU. Macro averaging, the average of the scores over the classes, was used to calculate the one value summaries for each metric. It was preferable to the alternative, i.e., micro averaging, since it assigns equal weight to all classes, contrary to assigning more importance to the higher-frequency class (Branco et al., 2016). As such, the macro summary will report low accuracy when there is low performance in the minor class, while the micro reports high performance overwhelmed by the majority class. Balanced accuracy is similar to macro averaged recall score. Rescaling the balance accuracy value between $1/(1 - number\ of\ classes)$ and 1, adjusted balanced accuracy was used to preliminarily identify models performing worse than a random classifier, i.e., score less than 0 (Brodersen et al., 2010; Guyon et al., 2015). On a single model assessment basis, the precision, recall, F1-score, and IoU for each class, along with the confusion matrix, were analyzed to distinguish which class is poorly classified.

Moreover, the precision-recall curve was utilized to analyze the uncertainty related to model predictions and further refine good-performing models (based on previous metrics) through the visual illustration of model precision and recall at different model confidence levels. This approach enabled the interpretation of model performance as a trade-off in information quality relating to the various purposes and target users of the results.

5.4. Experimental setup

The patch images dataset was split for training, validation, and testing (evaluation), as shown in Table 5.2.

Table 5.2 Train-validation-test data split of Sentinel-2 and Planet patches

Source Mosaic	Total number of patches	Training set (70%)	Validation set (15%)	Testing set (15%)
Sentinel-2	619	433	93	93
Planet	795	556	120	119

Experiments were set up based on the two machine learning models selected to assess their performance in the task of road pavement quality segmentation from optical satellite imagery datasets sourced and processed via different methods. The imagery datasets consist of the raw (spectral unmixing not applied) Sentinel-2 images (S), unmixed Sentinel-2 images (SU), combined Sentinel-2 images (SC) (unmixed stacked together with the raw ones), and Planet images (P). These datasets were analyzed using the U-Net and random forest models to establish eight experiments, as shown in Table 5.3.

Table 5.3 Model experiments undertaken and their designation

Model	Dataset	Experiment Designation
U-Net	Raw Sentinel-2 images (S)	DL-S
	Unmixied Sentinel-2 images (SU)	DL-SU
	Combined Sentinel-2 images (SC)	DL-SC
	Planet images (P)	DL-P
Random Forest	Raw Sentinel-2 images (S)	RF-S
	Unmixied Sentinel-2 images (SU)	RF-SU
	Combined Sentinel-2 images (SC)	RF-SC
	Planet images (P)	RF-P

All the experiments were conducted using a 12 core Intel® Core™ i7-9750H CPU @ 2.6 GHz with 16 GB RAM and a single NVIDIA Quadro T1000 GPU with 4 GB dedicated memory. The U-Net model is accelerated by the GPU (and partially by the CPU for processing data feed), while the CPU undertakes the processing for the RF model. Multiprocessing and multithreading approaches were applied where possible. All the code implementations in this research were based on python programming language using the library modules and frameworks described in Table 5.4.

Table 5.4 Programming frameworks and libraries used for the implementation of the experiments

Tasks	Python Library/Framework
Data acquisition	GEE python API, Planet python API
GMM clustering, PCA, RF model, and performance metrics	scikit-learn
U-Net model and data pipeline	TensorFlow
Feature creation	scikit-image
Model tuning	kerastuner
Raster and vector processing	Rasterio, GDAL/OGR

6. RESULTS AND DISCUSSION

This chapter reports and critically discusses the results of the experiments undertaken to investigate the potential of different strategies in using satellite imagery to extract road quality. Only the essential and most relevant results are displayed in this chapter to reduce cluttering. For the complete report of the results, see Appendix: Annex 7. The chapter will be divided into sections that transition from discussing the initial results to interpreting the sources of uncertainty and proposing modifications to address them.

6.1. Performance results of the experiments

6.1.1. Results of the RF model

The performance evaluation results of the RF model are presented in Table 6.1. The RF model produced an overall subpar performance across each accuracy metrics under all three datasets considered for the experiment. Experiment RF_SU, trained and tested on the unmixed dataset, showed the poorest performance with a recall accuracy of 44.88% (below 50%) and an even lower precision of 28.15%. This result indicates that the RF_SU model was not able to classify the RPQ labels sufficiently and correctly. The corresponding adjusted balanced accuracy, i.e., 28.15%, was also edging close to zero, indicating a near-random prediction. This aspect implies that the abundance maps from the unmixing were not helpful features for the RF model to classify road conditions. The lower performance of RF_SC compared to RF_S, indicating the degrading effect of adding the abundance maps as features, further demonstrates this aspect. Considering the correctness of predictions (how many predictions were actually correct), i.e., precision and an F1-score, the RF_S model showed the best performance with a precision of 34.03% and an F1-score of 37.67%. On the other hand, at the cost of incorrectness, the RF_P was able to return the most complete classification with a recall of 57.90%. The harsher metric, IoU, was commonly below par for all experiments, with the top performance coming from the RF_S experiment (32.22%).

Table 6.1 Comparison of the performance of different experiments considered for the RF model

Experiment	Macro Precision	Macro Recall	Macro F1-score	Macro IoU	Adjusted Balanced Accuracy
RF_S	**34.03%**	53.72%	**37.67%**	**32.22%**	38.29%
RF_SU	28.15%	44.88%	29.02%	24.96%	26.51%
RF_SC	32.13%	51.06%	35.22%	30.23%	34.74%
RF_P	29.81%	**57.90%**	31.60%	27.59%	**43.86%**

The confusion matrix for the best overall performing RF experiments, i.e., RF_S and RF_P, shown in Figure 6.1, will help further analyze the performance of the models (see Appendix: Annex 7 for the other confusion matrices). In each box intersecting between the true and predicted, the confusion matrix in Figure 6.1 shows the total classified pixels (top) and percentage correctly classified (i.e., prediction normalized) by the model (bottom). The color shades are visual guides, for which a strong diagonal (left to right) shade indicates good performance. However, as illustrated in both results, the box shades are stronger in the top two rows, i.e., 'non-road' and 'good' road classes. This aspect indicates that significantly large amounts of 'non-road' pixels are misclassified as 'good,' 'bad,' and 'very bad' road. The same can be observed between the 'good' road class vs. 'bad' and 'very bad' road classes. These misclassifications drag the precision and F1-score of the corresponding classes (see Table A 7.8), which ultimately drag the macro summation scores of the classifier to be very small. This misclassification is strongly evident in experiment RF_P (right), and even more so in the other two models, i.e., RF_SU and RF_SC (see Figure A 7.4 and Figure A 7.5).

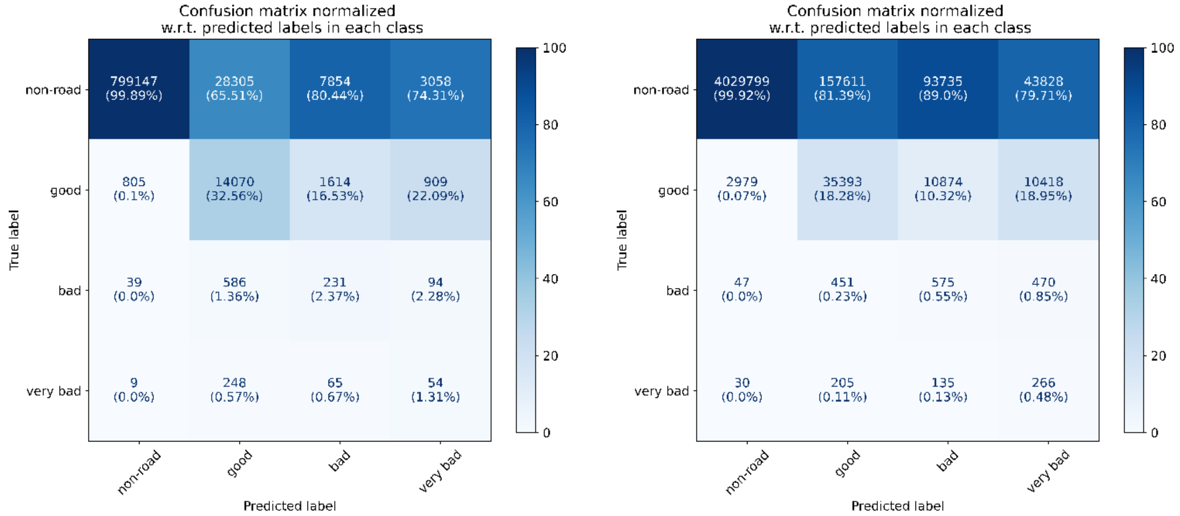


Figure 6.1 Prediction normalized confusion matrix of experiments RF_S (left) and RF_P (right)

The two basic criteria to assess the performance of the models are: (1) whether they can correctly classify the road pixels from non-road ones, and (2) the classification capability between good, bad, and very bad road segments. In this sense, the former criterion poses strong importance in distinguishing between asphalt (tar) road and bare soil or unpaved road. This distinction can be made in one of the two ways: the model can make hard classifications, labeling bare soil and unpaved road areas as 'non-road,' or it can classify it as 'bad road' or 'very bad road.' The latter is feasible since damaged roads display similar spectral characteristics as bare soil/unpaved (or completely worn out) roads (Herold et al., 2004, 2008). Moreover, under the scope of this study, it is of less importance to necessitate stringent delineation of road segment from other land covers (except bare soil and unpaved road) than it is to detect road distresses correctly and completely.

Accordingly, both models overpredict for the 'road' class in general (good, bad, and very bad), with the RF_P showing very strong signs of this misclassification, illustrated by the darker shades in the last three columns at the top row in Figure 6.1 (right). However, they both do predict most of the true road segments correctly, with a very small number of true 'road' pixels misclassified as non-road, as shown in the lighter shades in the bottom three rows of the first column of Figure 6.1. Within the three road condition classes, similar overestimation can be observed in both models; the RF_S model, however, shows more conservative predictions by classifying most road segments as 'good' (see Figure A 7.3).

6.1.2. Results of the DL model

The performance results of the U-Net model (DL), shown in Table 6.2, indicate a slight improvement from the RF model performance across all metrics, particularly the DL_S and DL_P experiments. The DL_SC model performed worse than its corresponding RF_SC experiment with an F1-score and IoU measure of 29.31% and 24.94% compared to 35.22% and 30.23% of the latter. The recall rates showed significant improvement for all the DL experiments. The relatively poor performance in using the SU and

SC datasets persists in the DL model, which imposes the need for further assessment of the sources of error in the unmixing procedure to explain the performance degradation effect of the obtained abundance maps as features for learning. Model DL_S, with a 37.93% F1-score and 32.40% IoU, can be selected as the best performing DL model with respect to all the measures except for recall, on which the DL_P model edges slightly with a score of 61.39%. The DL_S model outperforms all the RF models as well.

Table 6.2 Comparison of the performance of different experiments considered for the DL model

Experiment	Macro Precision	Macro Recall	Macro F1-score	Macro IoU	Adjusted Balanced Accuracy
DL_S	**35.62%**	61.32%	**37.93%**	**32.40%**	48.43%
DL_SU	28.69%	48.99%	29.32%	25.22%	31.99%
DL_SC	28.92%	52.88%	29.31%	24.94%	37.17%
DL_P	34.90%	**61.39%**	36.54%	31.75%	**48.51%**

The confusion matrix for experiments DL_S and DL_P are shown in Figure 6.2 (see Appendix: Annex 7 for the other confusion matrices). A similar trait of misclassifying ‘non-road’ pixels as ‘good’ or ‘bad’ or ‘very bad’ observed in the random forest models is also displayed in the confusion matrix of the DL_S and DL_P models. They show similar performance as the RF_S model in identifying the road pixels. DL_P, however, shows the most balanced road distinction (correctness and completeness). In distinguishing between the three road classes, both models show more or less similar performances. However, they have contrasting performance compared to RF_S, producing more complete RPQ segmentation at the cost of stronger overprediction of road defects (shown by higher percentages in the last two columns of the second row).

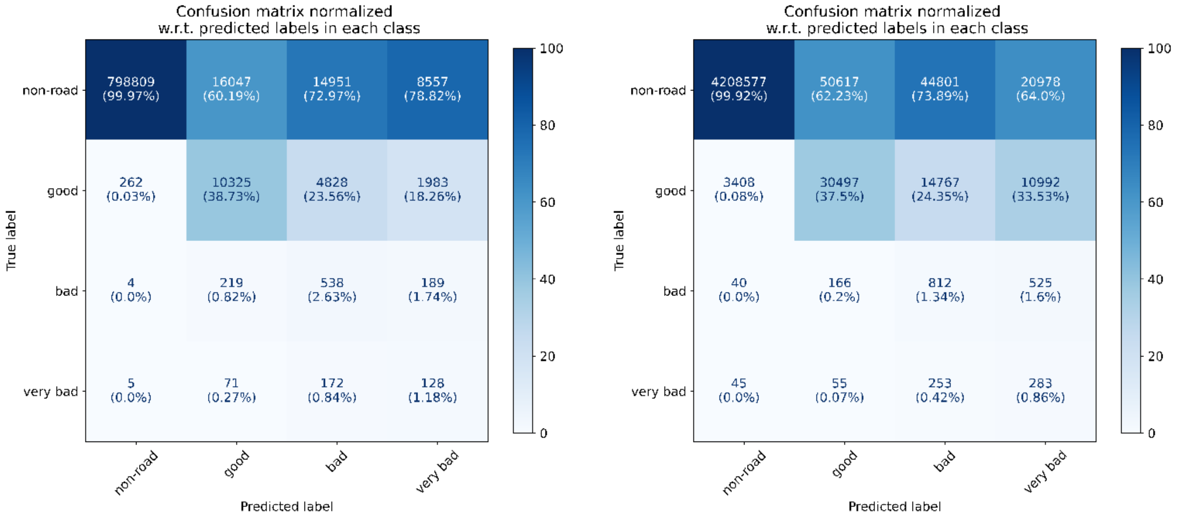


Figure 6.2 Prediction normalized confusion matrix of experiments DL_S and DL_P

6.1.3. Visual comparison of the results

An example of the prediction of the four models selected randomly is shown in Figure 6.3(a). The overestimation of the ‘bad’ and ‘very bad’ road classes indicated upon by the performance metrics is visually displayed here. Significantly larger segments of the road are predicted to be bad or very bad compared to the reference data. This aspect is more apparent in the models trained on Planet imagery Figure 6.3(b), i.e., RF_P and DL_P, in which almost all the road section is predicted as either in bad or very bad condition. This overestimation can be attributed to the fact that these models were tuned to maximize the weighted balanced accuracy (macro recall), which encourages the predictions of the minor classes, i.e., 'bad' and 'very bad' road in this case. A peculiar case is displayed in the result of RF_P in which the road condition labels are overestimated in contrasting areas to the reference. Though RF_S shows the least exaggeration of the poorness of road quality, it fails to extract the road section within the urban area. In this regard, the DL_P experiment shows a good distinction between road and urban coverage, while the other two, i.e., DL_S and RF_S, tend to misclassify some of the urban coverage as road or 'bad'/'very bad' class. The road is wider in all the experiment predictions, which explains the large number of misclassified 'non-road' labels.

Figure 6.4 demonstrates the effect of water features within the image on the predictions of the models. All experiments misclassify the river in the image as a 'good' road or even 'very bad' road. The reason for this effect is the spectral similarity of asphalt road and water in satellite imagery (Yang et al., 2018) and, more related to this case, the small representation of water features in the imagery dataset. Additionally, the figure demonstrates the severe misclassifications by the models, particularly evident in models trained and tested on the Planet imagery data, Figure 6.4(b).

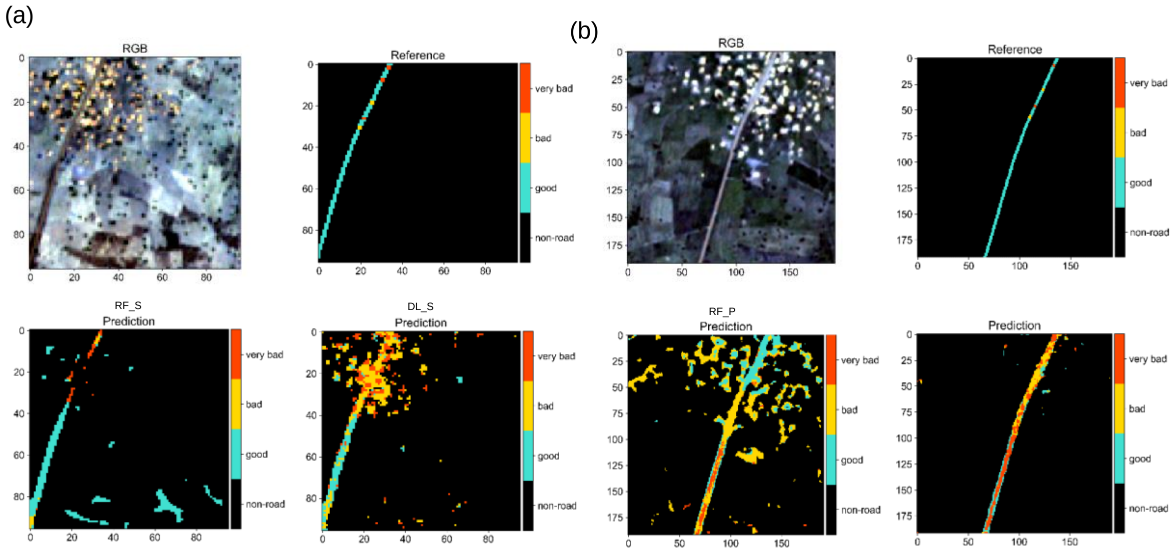


Figure 6.3 Comparison of predictions of (a) experiments RF_S & DL_S and (b) experiments RF_P & DL_P

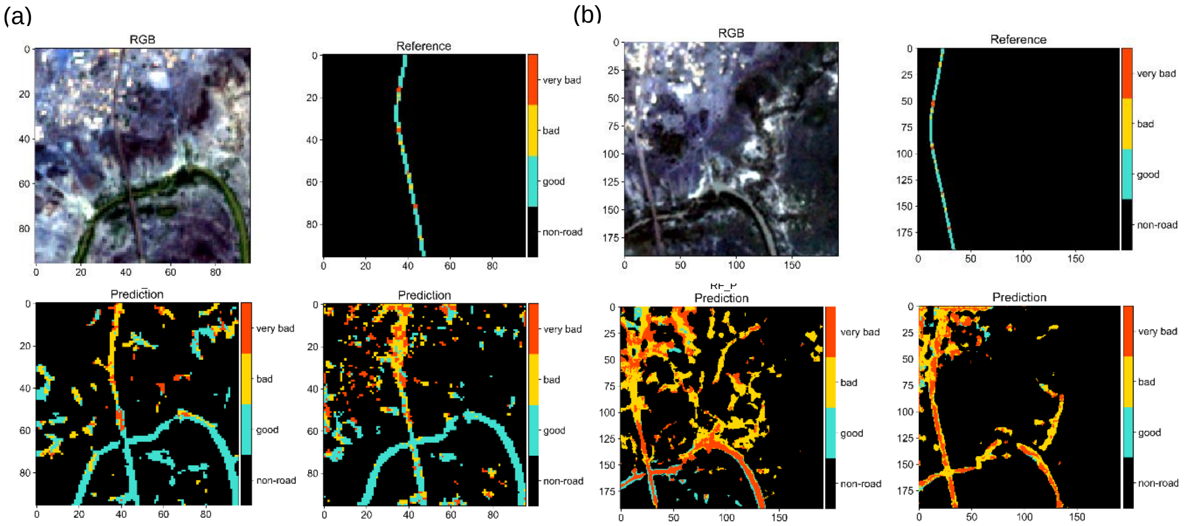


Figure 6.4 Comparison of the effect of water feature on the predictions of (a) models RF_S & DL_S and (b) models RF_P & DL_P

6.1.4. Concluding remarks on the results of the experiments

RQ-1.1. Does a deep learning method yield better performance in mapping RPQ compared to a shallow machine learning algorithm?

Overall, the top two performing experiments from each model were RF_S, RF_P, DL_S, and DL_P, as shown in Table 6.3. The DL_S experiment showed the best performance score, with DL_P and RF_S following at slightly lower scores. This aspect indicates that the deep learning method, in the form of U-Net, yielded better performance in mapping road pavement quality from optical satellite imagery rejecting the null hypothesis (H_0) under RQ-1.1. and confirming the alternative hypothesis (H_1). This result is expected considering deep learning models are particularly designed for computer vision tasks such as this. However, some factors on the implementations of the random forest, such as the mini-batch training and basic feature extraction, can be attributed to the lower performance of the RF experiments. Considering that their score falls only slightly lower from the top accuracy, improving those limiting factors and obtaining enhanced performance from the RF models is possible. Additionally, the RF models tend to predict RPQ more conservatively, favoring correctness more than complete detection, while the DL models favor completeness (recall) more than correctness. The use of focal loss, which encourages detection of minor classes, i.e., defects, can be attributed to this contrast.

Table 6.3 Summarized comparison of the performance of the two best experiments from each model

Experiment	Macro Precision	Macro Recall	Macro F1-score	Macro IoU	Adjusted Balanced Accuracy
RF_S	34.03%	53.72%	37.67%	32.22%	38.29%
RF_P	29.81%	57.90%	31.60%	27.59%	43.86%
DL_S	**35.62%**	61.32%	**37.93%**	**32.40%**	48.43%
DL_P	34.90%	**61.39%**	36.54%	31.75%	**48.51%**

RQ-2.1. How does using Planet imagery with higher spatial resolution affect the performance of the model compared to using Sentinel-2 imagery?

Comparing the datasets used in the experiment, the Sentinel-2 dataset, despite having a lower spatial resolution, produced better results in both models compared to the Planet imagery. This result confirms the alternative hypothesis (H_1) under RQ-2.1, which stated that using Planet imagery would yield equal or worse performance in extracting RPQ compared to using Sentinel-2 imagery. This outcome can be attributed to the additional two bands in Sentinel-2 imagery, i.e., *SWIR1 and SWIR2*, which are informative indicators of the condition of asphalt roads (Herold et al., 2004, 2008; Özdemir et al., 2020), that improved. Consequently, it demonstrates the viability of use Sentinel-2 imagery data to map road pavement quality. However, it is important to note that the results tend to overestimate the severity and extent of the defects compared to the reference data.

RQ-3.1. Can the performance of the models be improved through the application of spectral unmixing as a pre-processing step?

The poor performances exhibited by RF_SU and DL_SU indicate that the unmixing procedure does not improve the models' performance in mapping RPQ, which confirms the null hypothesis (H_0) under RQ-3.1. Combining the abundance maps (i.e., outputs from spectral unmixing) as additional bands with the raw Sentinel-2 images further reduced the models' performance compared to using only the raw ones. This aspect further consolidates that including spectral unmixing in preprocessing did not improve performance.

Although the results largely addressed the objectives outlined for this research, there is a substantial amount of uncertainty attached to these implications. In that regard, it is important to identify the points of uncertainty that hinder the potential of the proposed methodology, which the following subsection will address.

6.2. Sources of uncertainty

The challenge in assessing the quality of information about minute features such as road defects extracted from granular primary data lies in addressing the various levels of uncertainty that come along with it. The following sources of uncertainty were identified, considering the intricate details of processing undertaken during this investigation.

Imagery artifacts: refer to uncertainties arising from undesired features existing within the raw imagery data or developed during the basic pre-processing. It includes light (cirrus) cloud cover that might interfere with the spectral effect and cause misclassifications, mosaicking artifacts on the edge of original tiles shapes, and loss of information during resampling of Bands 11 and 12, among others.

Unmixing procedure: The unmixing procedure can be considered as one of the most error-prone elements in the methodology since it consisted of several intricate processing steps. Hence, the default in performance of the models using images processed through this method. Nevertheless, the value of this procedure in the task of road quality assessment is not to be diminished, hence entails further work to explore the applicable potential.

Reference data reliability: This source of uncertainty is particularly challenging because of the vector characteristics of the data causing complications in transforming to a label raster. In addition to the inherent spatial inaccuracy of the data, converting it to a label raster poses several challenges. All the vector transformations applied to locate the point within a tolerable level spatial inaccuracy (i.e., 50 m transverse snapping-to-road distance) combined with the effect of rasterization kernel can be pointed out as possible sources of uncertainty. Of course, the granularity of the Sentinel-2 imagery irons out these uncertainties to some extent. In this reasoning, the considerable degradation in model accuracy on the Planet data can also be attributed to this effect, among other causes. Nevertheless, a significant vector shift from the collected reference point can mean a difference in raster pixel being labeled as 'bad road,' and in this sense, the accuracy metrics can no longer be a fully viable option to decide on a performing model. Moreover, the reliability of the collected RPQ data, which can be significantly perturbated by various driver and vehicle behaviors (Wang, 2018), can cause significant challenges to the reliability of models trained based on them.

Differences in timeline: The various ranges of time with which the datasets for this investigation were collected have a substantial effect, particularly in this case, since road quality is not a static feature. In addition to the gradual changes in quality various transient visual interferences can occur over the road surface, such as sand cover and collecting water, that distort reflectances and induce 'false alarm' prediction of bad road section by the models. Although the latter features are more difficult to distinguish, their transient characteristics allow for smoother reports by aggregating several temporal predictions to even out their effects. Sentinel-2 imagery, with a 5-day cycle, is very well suited for this task. Nevertheless, it is still a factor of uncertainty to consider since such readings can mislead other more time-sensitive users, such as truck transporters planning their routes. On the other hand, the gradual changes made training and validating the models on the available data a difficult undertaking. Between the date of collection of the reference data and the sensing date of the imagery, the road pavement can deteriorate and display distresses or undergo rehabilitation such as patchwork. These changes would entail a discrepancy between the reference data and the information extracted from the satellite imagery and devalue the validation to some extent. In this sense, it would be difficult to tell which data is correct.

Undetectable defects: In realizing the limitations of the approach employed here, it is important to note that some road defects, perhaps at the threshold of 3 meters size as Radoux et al. (2016) suggested, cannot be detected in the Sentinel-2 imagery or even in Planet imagery (at 4.7 m resolution). However, the reference data is less limited by these traits and therefore includes some of these features, which by training the models on this data to detect these features, we are asking a near-impossible task. The resulting product can be deceptive by mimicking the reference data through the overfitting characteristics of ML models. Although the data has been preprocessed to remove some of these features based on their RPQ annotations, as mentioned in Section 5.1.2, it is possible that some might have passed through since their annotations do not describe their overhead distinguishability.

The fine line between 'bad' and 'very bad' roads: The categorical classification of road quality is often made subjectively. Since the standard of categorizing the severity of the detected defects of the reference data has not been reported, it has to be considered that these categories can be interpreted in other ways. Moreover, in using medium resolution satellite imagery to detect these features, the distinguishing line between these categories is further thinned since typically, they will only display minimal differences in spectral characteristics unless viewed with high-resolution HSI imagery (Herold et al., 2008). This effect can influence the performance of the ML models by inducing confusion between 'bad' and 'very bad' road classes, which is observable from the confusion matrix results in Sections 6.1.

Model uncertainty: ML classifiers are probabilistic learners, which means that they undertake classification by estimating the probability of a certain data belonging to a certain class. Accordingly, these classifiers have their own confidence level in assigning classes (Brownlee, 2020), which contributes to the overall uncertainty of the obtained information product.

Unknown utility: Considering the target users this research aims to address, i.e., transporters, traders, and road authorities in Ghana and Burkina Faso, it is implausible to decide how such products can be used without having a certain level of stakeholder engagement. Accordingly, it is valuable to realize that some users prefer products with low false alarm rates. In contrast, some would rather have high detection rates towards estimating the maximum risk incurred from these defects. Therefore, this unknown utility encourages the inclusion of some degree of flexibility in the model's predictions, a trait that will be further discussed in Section 6.2.5.

Towards addressing these uncertainties, the following subsections describe findings corresponding to the sources of uncertainty mentioned above.

6.2.1. Validation on Street View information

Street View images collected from Google Earth were used in this study with the aim of offering an additional degree of observation and reducing (if not accounting for) uncertainties caused by time differences, reference data reliability issues, and undetectable road defects as noted in the previous section. Fortunately, the Street View images available for long segments of the corridor were collected in April 2016, which is close enough to the reference observation timeline to infer some insight on the results obtained in this study. The results from the best performing model, i.e., DL_S, were used for this assessment. The figures presented below show the model prediction (top left), reference label (top middle), Google Earth image (top right), and the Street View images referenced by the numbered circles numbers in the top images.

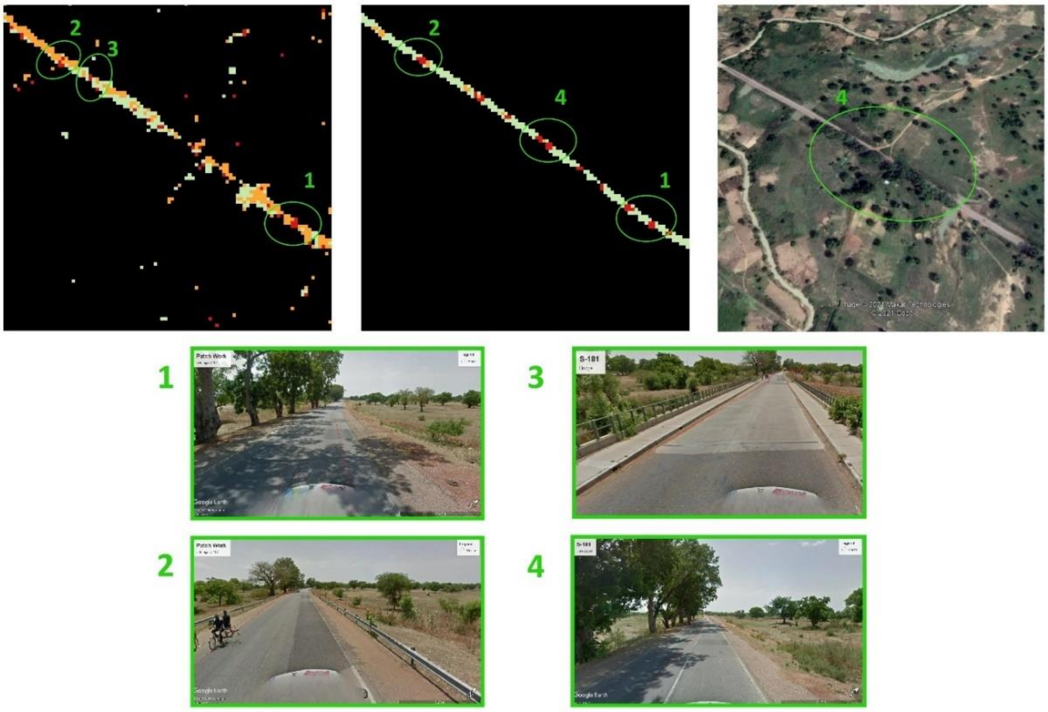


Figure 6.5 Model DL_S predictions compared with the referenced data and Street View images

From the first impression of Figure 6.5, a simple conclusion can be drawn that the model predictions are misclassifying most of the road section as 'bad' (indicated by large yellow coverage on the predictions). Upon observation of scenes from Street View, the road is covered with frequent patch fixes (at almost regular intervals throughout) that appeared to be applied close to the time of the Street View images. Noting from the RPQ annotations from the CHEETAH data, most of the RPQs in this section were rough patches. The circles labeled 1 and 2 are observations of these fresh patches, which were identified as a defect in the corresponding timelines of the reference data and predictions. The observation of these rough patches on Street View provides a plausible explanation that these patches deteriorated during the time difference and got detected as RPQ defects in the reference data and hence predicted as such by the models. The larger spatial extent observed as overestimation in the model prediction can be explained by the degree of deterioration of these patches at the time of imagery and the length of the patches, which, as shown in label 2 in Figure 6.5, can extend very long. A particular discrepancy observed in label 4, which is the missing road section in the predictions, is explained as a good performance in clipping out vegetation from the road while also showing predictions in between the occlusions. Moreover, label 3 shows a 'bad' road prediction also evident in Street View, missing in the reference data. The corresponding defects were found to be bridge expansion joints that were mislocated in the reference data upon inference from the RPQ annotations.

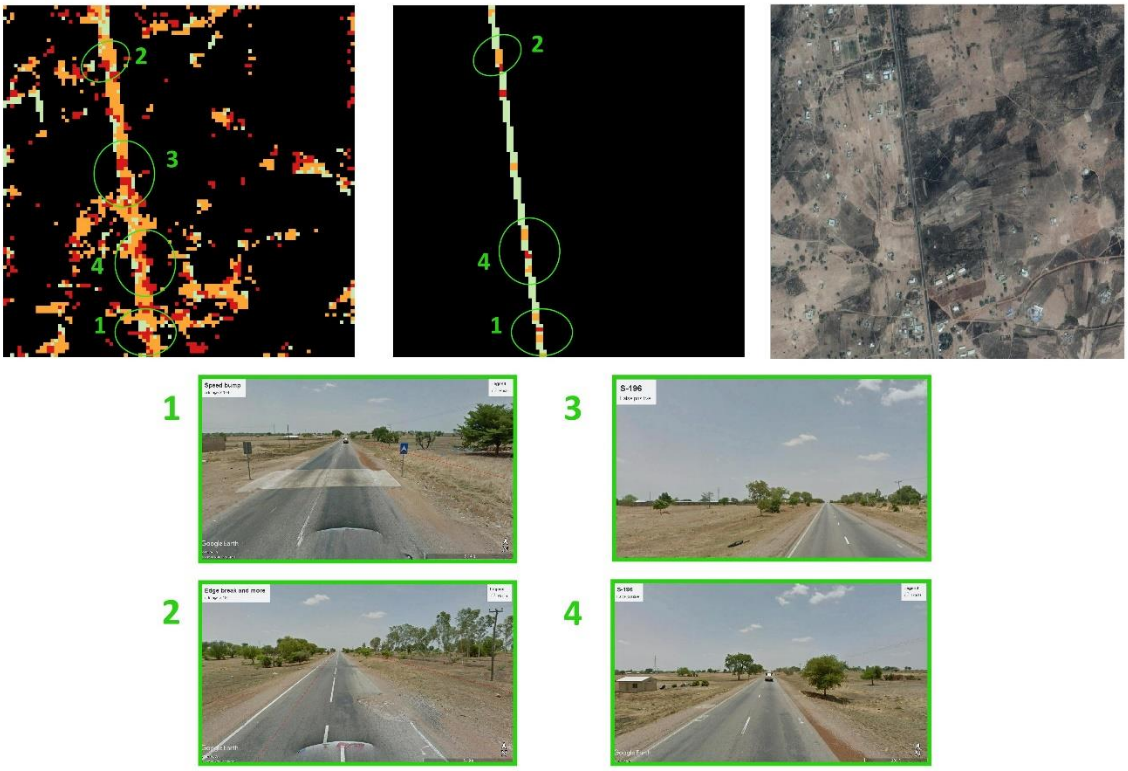


Figure 6.6 Model DL_S predictions compared with the referenced data and Street View images

Like the previous example, we can observe a large overestimation of road defects. In fact, almost all of the road section is classified as either 'bad' or 'very bad' road by the DL_S model. These exaggerated predictions were proven incorrect from observation of Street View on the road section, although the timeline and other factors can still be plausible explanations aside from model incorrectness. Circles 3 shows an instance of this misclassification where the model predicted a 'very bad' road, while no road defects were identified on the Street View images and the reference data displayed only a small patch of 'bad' road. On the other hand, circles 1, 2, and 4 demonstrate agreement between model prediction and reference data. Circle 1 and circle 2 were also confirmed on Street View, displaying a worn-out concrete speed table and an edge break-off, respectively. However, detection of road defects (annotated as speed bumps and rough patches) shown in the predictions and reference data under Circle 4 could not be confirmed on Street View.

6.2.2. Identifying undetectable defects

As noted previously, it is important to delineate the limits of the proposed methodology such that its utility is well defined. By comparing the predictions with the reference data and associating this information with the defect annotations that were also rasterized for this assessment, a simple statistic on the type of road defects correctly identified by model DL_S was obtained. This statistic is presented in Figure 6.7. From these findings, we could infer that some road defects, such as road joints, have poor detection rates while rumble stripes, which typically have broader spatial coverage, have more than 50% detection rate. Interestingly, out of the three corrugated road features, all were detected by the model. Although this assessment is very sparse to offer insight into selecting which defects to detect in such a methodology, robust information on these detectability rates can be estimated through the aggregation of dense temporal data collections. In this sense, it could enable the standardization of such remote sensing approaches in detecting road defects.

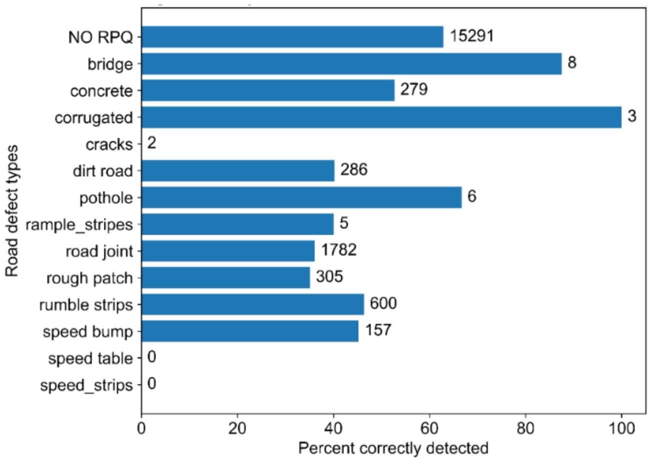


Figure 6.7 Percent correctly detected for each road defects

6.2.3. Three-class RPQ segmentation

At the cost of losing information depth, the classes of the RPQ segmentation were decreased from 4 to 3 (aggregating the 'bad' and 'very bad' classes as one 'bad' road class) to investigate the effect on mitigating the uncertainty related to the classification confusion between the two aggregated classes. Of course, this would entail an increase in the performance metrics scores, the significance of which will be investigated in this section. The DL (U-Net) and RF models were considered for this experiment undertaken with the plain Sentinel-2 and Planet imagery datasets. The designations for the experiments are RF_S_3C, RF_P_3C, DL_S_3C, and DL_P_3C, representing the RF model on the Sentinel-2 and Planet data, and the DL model on the Sentinel-2 and Planet data, respectively.

Accordingly, Table 6.4 shows the performance evaluation results of these experiments. In each experiment, significant improvements can be observed throughout all scores. Their adjusted balanced accuracies are also well above 50%, indicating an overall good performance. Experiments RF_S_3C and DL_S_3C were the top performers from their respective model groups (i.e., RF and DL). DL_S_3C tops the list in all measures with an F1-score and IoU of 53.65% and 46.03%, comparable to previous remote sensing-based road quality assessment works (Karimzadeh & Matsuoka, 2021) and even to recent road extraction works (Oehmcke et al., 2016). Figure 6.8 and Figure 6.9 show some example patches demonstrating the results of the three-class RPQ segmentation for experiments DL_S_3C and DL_P_3C. From these figures, we can observe the improvement in the overestimation of road defects, although it is still present.

Table 6.4 Performance results of the three-class RF and DL models

Experiment	**Macro Precision**	**Macro Recall**	**Macro F1-score**	**Macro IoU**	**Adjusted Balanced Accuracy**
RF_S_3C	45.33%	73.84%	48.78%	41.88%	60.77%
RF_P_3C	40.50%	75.80%	43.56%	37.86%	63.71%
DL_S_3C	**50.95%**	**81.44%**	**53.65%**	**46.03%**	**72.17%**
DL_P_3C	47.00%	75.89%	51.34%	44.36%	63.84%

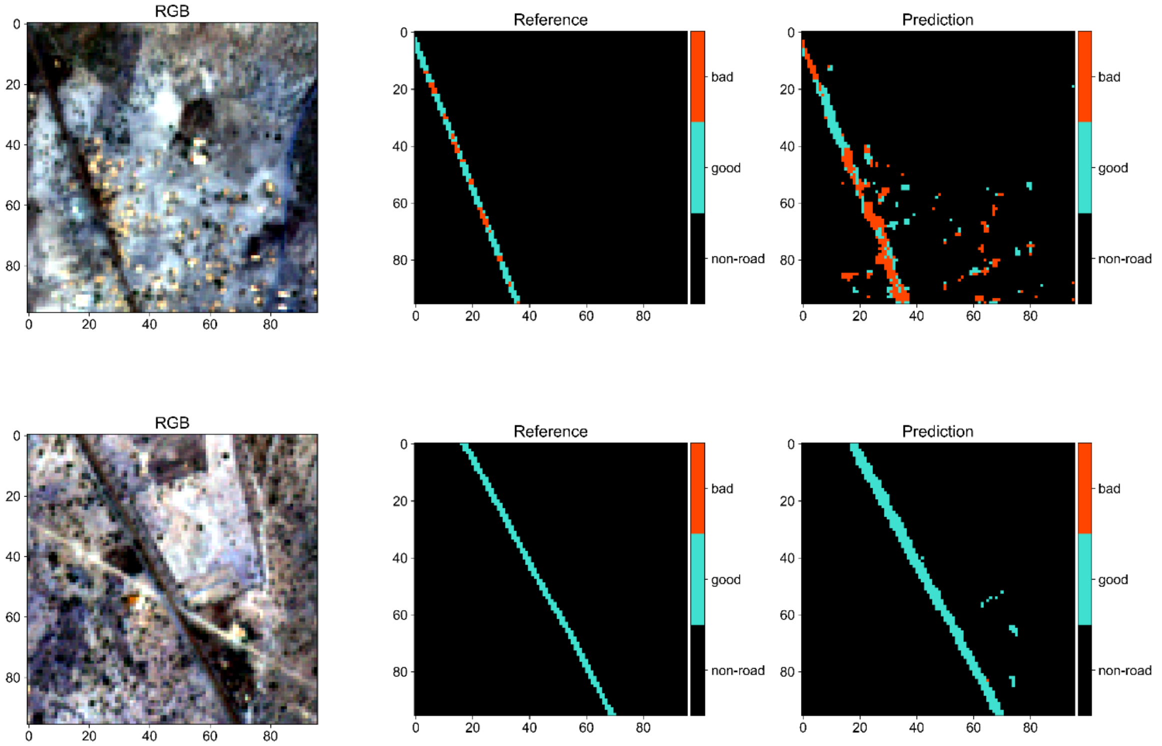


Figure 6.8 Three class RPQ segmentation result for experiment DL_S_3C

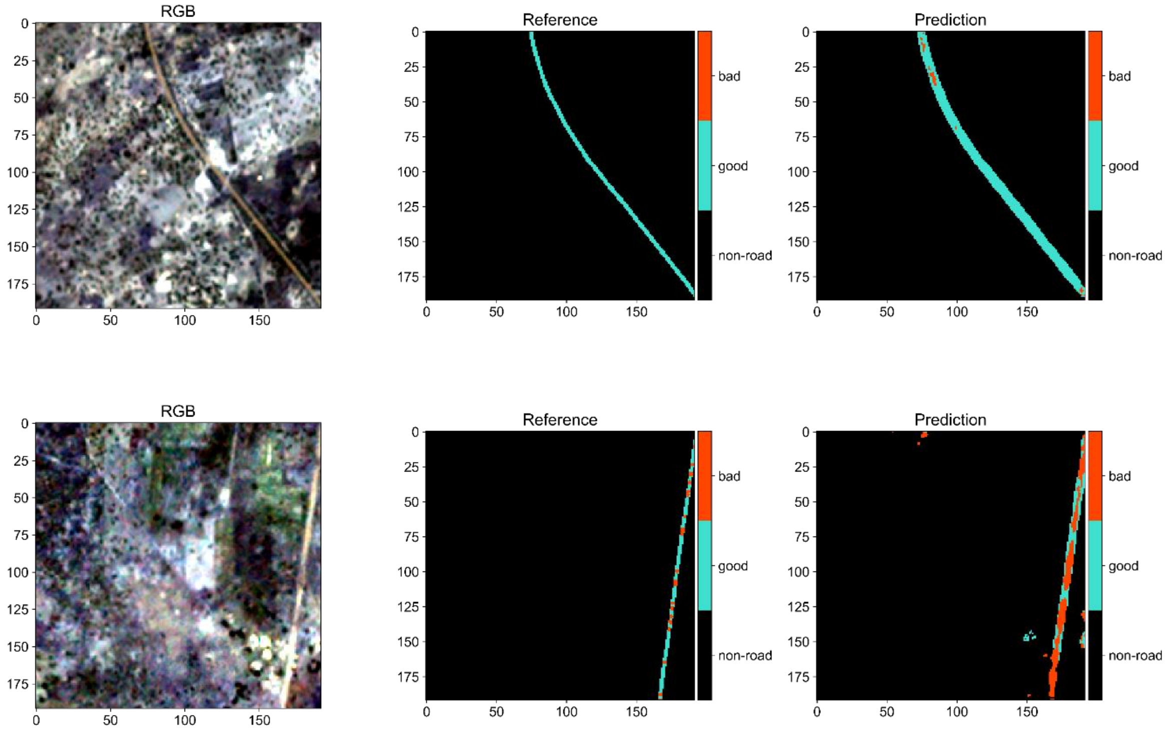


Figure 6.9 Three class RPQ segmentation result for experiment DL_P_3C

6.2.4. Spectral Analysis

Due to various pre-processing levels involved in this study's methodology, it is important to track how much spectral information is retained from the original features of RPQ defects to assess the source of uncertainty related to these pre-processing steps. This subsection discusses this aspect by describing the spectral characteristics of the datasets used at various processing stages. A basic spectral index-based analysis was used to describe these spectral characteristics. Accordingly, three spectral indices and their ideal land use/land cover classification indices adapted from Chen et al. (2006) were used. The indices used are Normalized Difference Vegetation Index (NDVI) (Rouse et al., 1973) in Eq.(6-1), Normalized Difference Water Index (NDWI) (B. C. Gao, 1996) in Eq. (6-2), and Normalized Difference Built-up Index (NDBI) (Zha et al., 2003) in Eq. (6-3). These indices are formulated to identify vegetation, water, and built-up areas, respectively. The recommended classification value ranges for different land use/cover types as used in Chen et al. (2006) are described in Table 6.5. It is important to note that the use of these indices is not intended as an additional hard classification method but as means of inferring distinguishing patterns of features in the data.

$$NDVI = \frac{NIR - RED}{NIR + RED} = \frac{Band\ 8 - Band\ 4}{Band\ 8 + Band\ 4} \tag{6-1}$$

$$NDWI = \frac{GREEN - NIR}{GREEN + NIR} = \frac{Band\ 3 - Band\ 8}{Band\ 3 + Band\ 8} \tag{6-2}$$

$$NDBI = \frac{SWIR - NIR}{SWIR + NIR} = \frac{Band\ 11 - Band\ 8}{Band\ 11 + Band\ 8} \tag{6-3}$$

Where GREEN, RED, NIR, and SWIR represent the green, red, near-infrared, and short-wave infrared channels of the spectrum. The equations on the left-hand side are those used as per Sentinel-2 band notation.

Table 6.5 Spectral indices value ranges for different land use/cover types adapted from Chen et al. (2006)

Land use/cover types	NDVI	NDWI	NDBI
Built-up	< 0.2	< 0	0.10 – 0.30
Bare land	< 0.2	< 0	> 0.25
Vegetation	> 0.2	> 0.05	< 0
Water	< 0	> 0	< 0
Semi-bare land	> 0.2	> 0	> 0.2

At first, we will assess the spectral characteristics of pixels categorized as ‘good,’ ‘bad,’ and ‘very bad’ road based on the referenced data. Box plots were used to better visualize the distribution of the data across the values of the indices. Figure 6.10 shows the box plots of NDVI (top left), NDWI (top right), and NDBI (bottom center) for each class. From this figure, we can infer that most of the pixels in each class have NDVI values below the threshold (indicated by the top of the boxes, i.e., the upper quartile or at least

75% of the data, falling below). This aspect means that these classes are considerably differentiable from vegetation. The plot also indicates the consistency of these values, as shown by the small height of the box and whiskers. However, some representative data points (around 25% for the 'good' class) and the outliers display higher NDVI values. This aspect is typically an indication of mixed pixels caused by nearby or over-the-top (tree) vegetation imposing spectral interference on the road. The box plot of the NDWI shows a clearer distinction between water features and roads, with only some outliers that can be explained by the inherent spectral similarity between fresh road and water features. The NDBI plot is trickier to interpret because the values for each class show slight differences, and the data points in each class are centered within the range of the built-up area indicator. Moreover, many outliers fall below that range, possibly due to the mixed pixel effect, as noted earlier. However, a slight rising trend is observable in the center of the data points, moving from 'good' to 'very bad,' indicating a potential distinguishability between the classes.

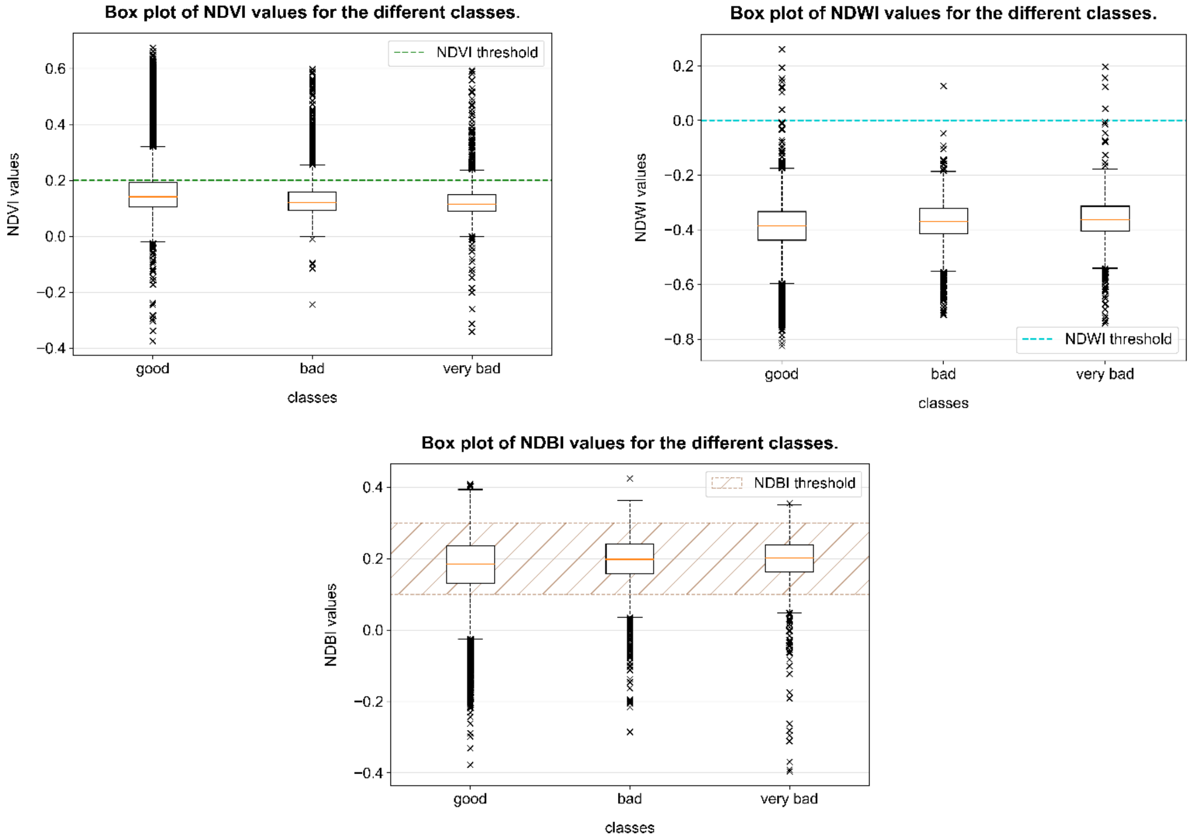


Figure 6.10 Box plots of NDVI (top left), NDWI (top right), and NDBI (bottom) values of each RPQ class

The spectral characteristics of endmembers obtained during unmixing can offer good insight into the drastic degradation caused by including these features for the ML algorithms. Primarily this helps in characterizing the clusters and determining whether there is a significant distinction between them. It also enables the proposal of mitigations for further research in this field. Accordingly, Figure 6.11 shows the respective box plots for the endmember clusters found by the GMM method employed in this study. From the NDVI boxplot, it can be inferred that cluster 1 (falling completely above the threshold) refers to vegetation, and cluster 5 can be an effect of mixed pixels or a semi-bare land cover. The NDWI plot distinguishes cluster 2 to contain some water feature endmembers as outliers. A possible explanation of this aspect is that cluster 2 refers to noisy data points that do not have enough representation to form clusters, which in this case would be water features and rooftops. The center of this cluster was shown to fall in the range of built-up coverage on the NDBI boxplot, further consolidating the explanation above. Cluster 3 stands out in the NDBI plot by large portions of the datapoints displaying high NDBI values followed by clusters 4 and 5 (excluding cluster 2). Consequently, given their significant spectral separability, these three clusters can reasonably be considered features usable to distinguish between 'good,' 'bad,' and 'very bad' road and bare soil.

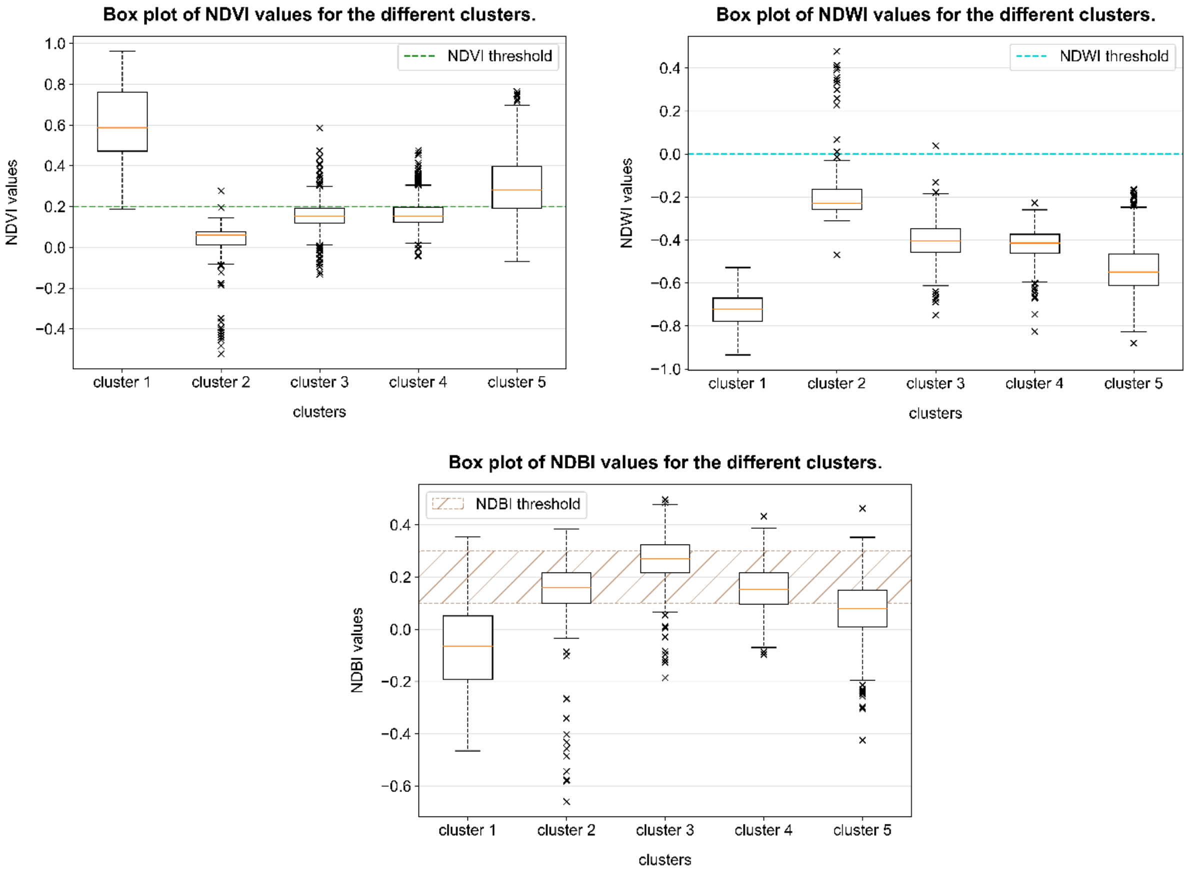


Figure 6.11 Box plots of NDVI (top left), NDWI (top right), and NDBI (bottom) values of each endmember cluster

With these preliminary suppositions, it is evident that there is significant value in applying unmixing as a pre-processing technique. Nonetheless, the difficult task is transforming these endmembers into features usable by the ML algorithms, i.e., abundance maps. This transformation is strongly sensitive to the distribution of the data in each cluster. If there are outliers representing bare soil within a cluster describing 'good' roads, then it would mean that while most feature maps of this cluster represent the former, some will refer to the latter. This aspect confuses ML algorithms which depend on the consistency of the structure of features in the data, thereby degrading the learning process. Unfortunately, this defect was evident in the abundance maps obtained after unmixing (see Figure A 7.7). The creation of blank bands further exacerbates this aspect due to the absence of an endmember that belongs to that cluster representing the band within the patch considered.

Finally, to assess the spectral features of the RPQ predictions, the results of experiment DL_S were explored (since it is the best performing model and enables the calculation of NDBI). Accordingly, the boxplots shown in Figure 6.12 describe the spectral characteristics of the RPQ classes inferred by the DL_S model. The plots show more or less similar traits as the reference RPQ pixels, albeit with significantly larger outliers in each class for all three indices. Looking at the NDBI in particular, we can observe that the distribution center of the ‘good’ class falls lower than the other two, which entails ‘bad’ and ‘very bad’ class roads appear more similar to bare soil than ‘good’ roads do. This trait is also slightly apparent in the reference data, as noted previously. Although these traits are good indications of the ML model’s learning effectiveness, they are merely data statistics. The extending 25 percentile and outlier datapoints of each class would indicate significant misclassifications and thereby introduce more uncertainty to the results.

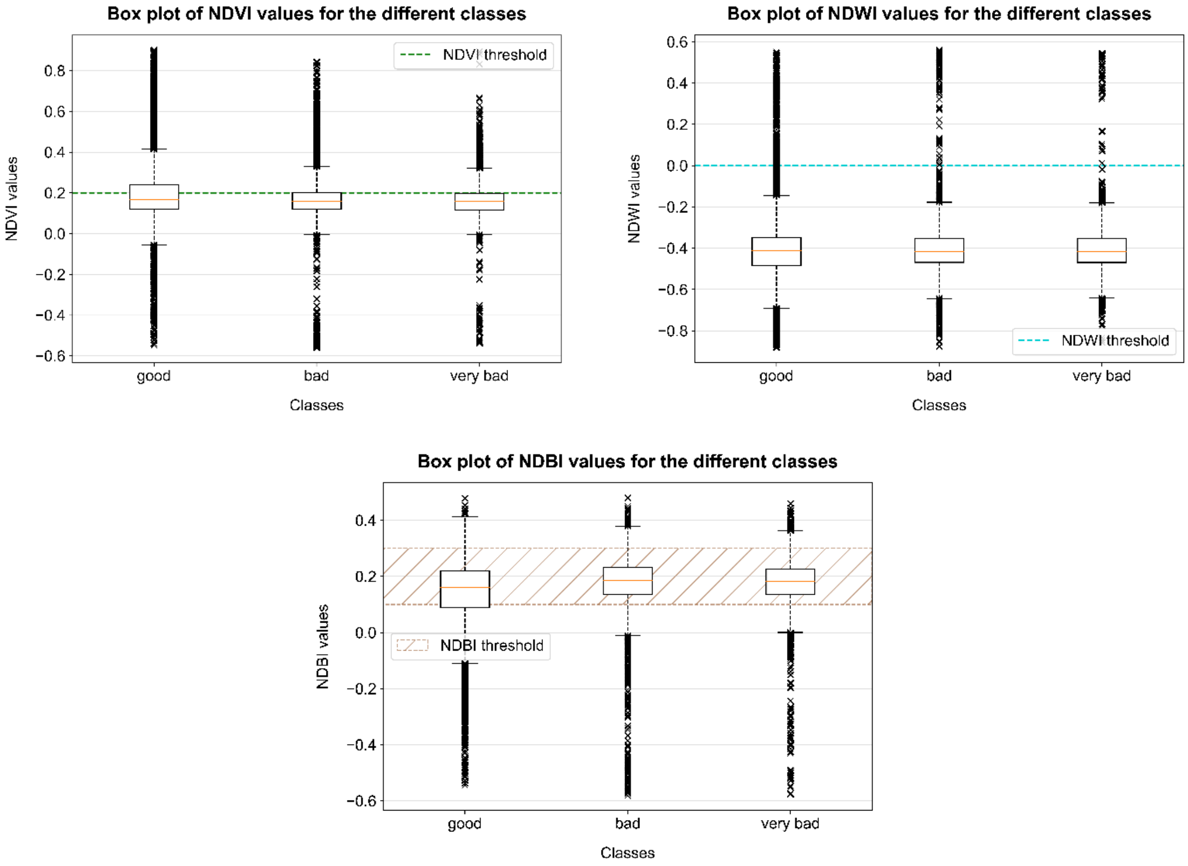


Figure 6.12 Box plots of NDVI (top left), NDWI (top right), and NDBI (bottom) values of each predicted class

6.2.5. Prediction probability analysis

Supervised ML classifiers, including those used in this study, are typically probabilistic models (exceptions include Support Vector Machines). Probabilistic models make classifications by assigning probabilities to each class and naturally applying thresholds equivalent to the inverse of the number of classes, i.e., 0.5 for binary classifications, to produce crisp classification labels (Fernández et al., 2018). In this sense, for a binary classifier, if a class has a probability value of more than 0.5 for a data point, then that data point will be assigned to that class. The probabilities provide information on the model’s confidence in assigning that class and therefore are important factors to consider in assessing the uncertainty related to ML model classification. In practice, the thresholds for classification can be selected based on the application intended (Brownlee, 2020), which is valuable in this case to understand the uncertainties associated with model choices and propose trade-offs based on the uncertainty tolerance of the users. More importantly, the natural thresholds of ML models often may not represent the optimal classification based on the probabilities due to class distribution imbalance and the biased cost of misclassifying one class more than the others (Fernández et al., 2018). The former cause is strongly evident in our case, while the latter is also an important factor in deciding the information quality as per the targeted user. According to this interpretation, the choice of the threshold depends on the optimization of model performance as well as the trade-off between the desired class error and the tolerable counter class error. This threshold selection is called threshold-moving and is commonly applied through the trade-off between precision and recall over the precision-recall curve (pr curve). The precision-recall curve is particularly suitable for an imbalanced dataset as it is intuitive in interpreting model prediction on imbalanced data (Saito & Rehmsmeier, 2015). Through threshold-moving, it is possible to evaluate a model using the pr curve, interpret the trade-offs, select the optimal threshold, and apply the threshold to make new predictions (Brownlee, 2020).

In applying this approach for this assessment, the best performing model, i.e., DL_S, was selected here again. Figure 6.13 presents the frequency of prediction probability (i.e., how many times the model assigned a probability value to a certain class) for the selected model. This figure informs us of how confident the model assumes it is in predicting the respective class. In that regard, while the non-road class (top left) frequently shows high probabilities, the other classes are showing very small (0.0 to 0.2) probability values in most of their predictions. This aspect comes as an effect of class imbalance. Moreover, the probability maps obtained by removing the softmax operation from the last layer of the U-Net model are shown in Figure A 7.8. They display the spatial distribution of the probabilities assigned for each class along with the final predictions made. From the figures, we can infer how small the threshold for the 'bad' and 'very bad' classes are set to be, which can explain the priorly observed model overestimation of those classes.

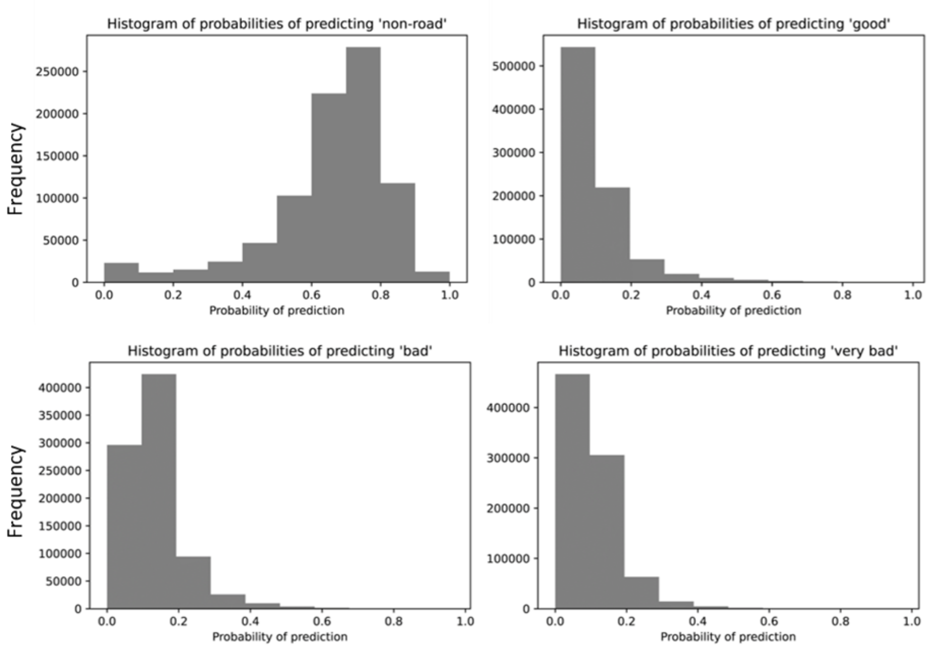


Figure 6.13 Frequency histogram of the prediction probabilities for non-road (top left), good (top right), bad (bottom left), and very bad road (bottom right) class

Based on the findings above, the precision-recall curve (pr curve) was employed to select the optimal threshold for the DL_S model prediction. Figure 6.14 presents the pr curve produced for the model (see Appendix for the pr curve of model DL_P for performance comparison). The dashed horizontal lines calculated from the frequency of the class in the dataset are an indicator for random performance, i.e., the same as the 0.5 probability for predicting a coin toss. A curve that falls below that line is considered less performant than a random classifier. This line is very high for the 'non-road' class (black dashed line), while it is barely visible for the 'bad' and 'very bad' classes. This aspect is an indication of the heavy class imbalance within the data. Based on the previously described criteria, the classifier for 'non-road,' 'good,' and 'bad' road classes are shown to perform better than a random classifier (black, turquoise, and yellow solid lines, respectively). However, the classifier for 'very bad' road (solid red line) falls very close to the random classifier line, and it is possible that at some threshold value, it can fall below it. Considering the interpretation of these curves (Saito & Rehmsmeier, 2015), the 'non-road' classifications display a near-perfect performance, while that of the 'good' road class are considered good performances. Given the degree of imbalance, the classifications of 'bad' roads can be taken as fair classifiers. However, the 'very bad' road classifier indicates poor performance, although not random in totality. The area under the curves (AUC) can also be used as a performance metric, with similar characteristics as F1-score.

The selection of optimal threshold can be made from this curve by maximizing the F1-score, which is a balancing average between precision and recall and therefore optimizes both simultaneously. The precision-recall coordinates that correspond to this threshold are shown in the figure by the bold points on individual curves with the optimizing threshold and the respective maximum F1-score indicated in the plot legend. The class-wise selected thresholds, corresponding scores, and the overall performance are summarized in Table 6.6. Alternatively, these points can be visually located by using the iso-f1 curves, which are lines connecting points in the pr curves with equal F1-score.

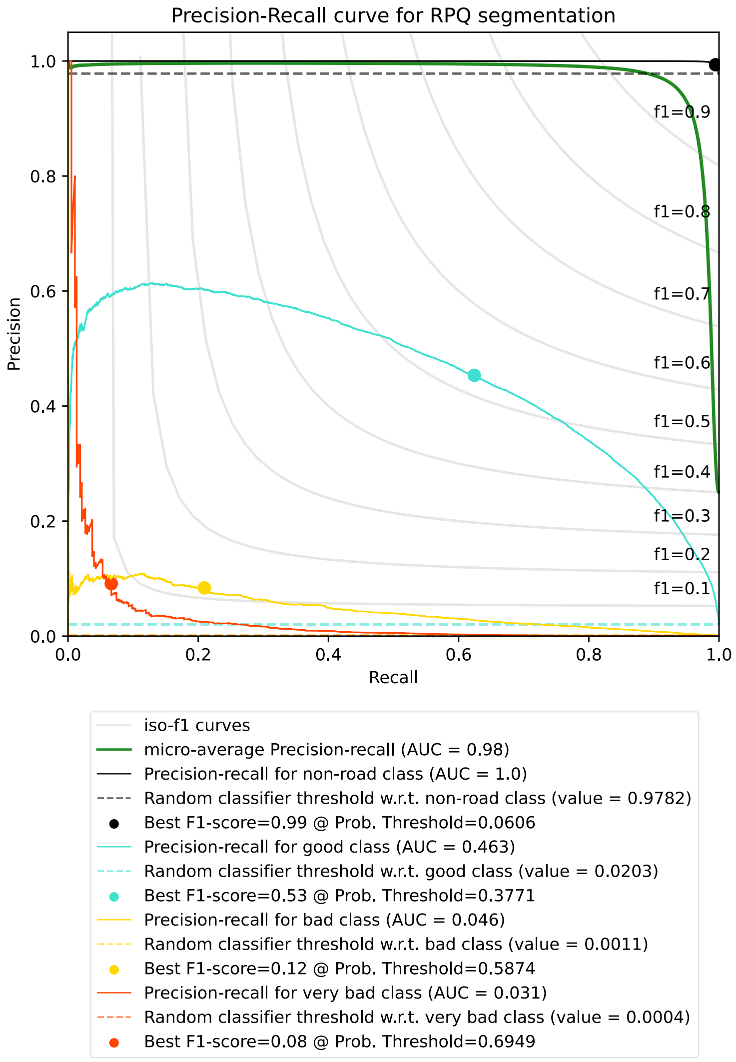


Figure 6.14 Precision-recall for curve for model DL_S

Table 6.6 New performance scores of DL_S based F1-score optimized thresholding

		Metrics			
Class	**Threshold**	**P**	**R**	**F1**	**IoU**
Overall		**40.80%**	**46.80%**	**42.89%**	**36.07%**
Non-road	0.06	99.24%	98.45%	98.84%	97.71%
Good	0.38	46.98%	61.71%	53.35%	36.38%
Bad	0.59	8.37%	20.42%	11.87%	6.31%
Very bad	0.69	8.59%	6.65%	7.50%	3.89%

The above thresholding leans to the conservative side, meaning that it favors more confidence (especially in the RPQ classes) than it does detections. For a sense of comprehensive understanding, the following tables present alternate thresholding results: the first one is based on an empirically selected threshold for this study, and the second one is from the opposite end of a conservative choice, i.e., favoring more detection (see below). The results demonstrate the corresponding trade-offs between precision and recall.

Table 6.7 Performance scores of DL_S for empirical and non-conservative thresholding

			Metrics			
Choice	**Class**	**Threshold**	**P**	**R**	**F1**	**IoU**
Empirical	**Overall**		**35.95%**	**60.67%**	**39.33%**	**33.41%**
	Non-road	-	99.93%	96.15%	98.01%	96.09%
	Good	0.28	37.84%	66.75%	48.30%	31.84%
	Bad	0.46	4.78%	40.95%	8.56%	4.47%
	Very bad	0.28	1.26%	38.83%	2.44%	1.23%
Non-conservative	**Overall**		**35.62%**	**61.32%**	**37.93%**	**32.40%**
	Non-road	-	99.97%	95.28%	97.57%	95.25%
	Good	-	38.73%	59.35%	46.87%	30.61%
	Bad	0.1	2.63%	56.63%	5.02%	2.57%
	Very bad	0.1	1.18%	34.04%	2.28%	1.15%

6.3. Limitations of this study

This research recognizes two significant limitations in the undertaking of the investigation. The first limitation is the lack of reliable reference data for validation. Although the CHEETAH RPQ dataset was used as a reference to validate the experiments, the complications identified in the previous subsections raised challenges in validating the results of this study. An ideal source of reference data for validation, in this case, would have been in-situ collected data. However, due to the retrospective nature of this study, in-situ data was not possible to collect. Therefore, it would be of significant value to reiterate this work using ground collected data as a reference towards suitably evaluating the potential of the proposed approach.

Secondly, due to time limitations, the experiments designed for this study were not tested for variance in model initialization (i.e., different data split, neural network weight initializations, etc.) via multiple trials. These variances can be an additional source of uncertainty related to the models that can influence the applicability of the models. Moreover, the time limitation also restrained tuning of the models such that the optimal hyperparameter combinations were selected based on the Sentinel-2 dataset. Although the U-Net model was slightly adjusted for the Planet imagery dataset by reducing the mini-batch size and learning, it is possible to obtain better results from the Planet dataset if the models were tuned accordingly.

7. CONCLUSION AND RECOMMENDATIONS

With the motivation of addressing post-harvest losses (PHLs) in developing countries, this study identified the need for spatial information on road pavement quality among these regions. On the basis of a multi-level conceptual framework for causes of food loss (HLPE, 2014), it was recognized that the provision of this information at a lower quality but in a timely and accessible manner could mitigate PHLs that occur due to poor road infrastructure.

In this regard, freely accessible and lower resolution optical satellite imagery hold significant potential as primary resources to extract road pavement quality information. However, perhaps due to the complexity of the task, there was limited existing research regarding the extraction of road quality information from lower resolution optical satellite imagery, such as Sentinel-2. Upon establishing the lack of a methodological body of work in this field, this research investigated the use of machine learning methods in the considered task. Accordingly, two machine learning (ML) models: a random forest and a U-Net model, representing the two frameworks of ML, i.e., shallow and deep learning, were explored. Additionally, a novel approach of integrating a spectral unmixing procedure was proposed to address the challenge of identifying minute details from low-resolution imagery. This unmixing procedure involved a local unmixing strategy that made use of the patch-based learning of the ML models to apply localized unmixing using a sparse statistical algorithm called SPICE (Zare & Gader, 2007). Towards this end, experiments were undertaken to investigate the performance of the proposed methods in using Sentinel-2 satellite imagery to extract road pavement quality over a road corridor based on reference data collected for the road through crowdsensing.

The results, although objectively subpar, were promising in realizing the use of such data sources for the timely accessibility of road pavement quality information. The best model, i.e., the U-Net deep learning architecture, reported an F1-score of 37.93% and an IoU of 32.40%. This result indicates that the deep learning model, i.e., U-Net, outperforms the shallow ML algorithm, i.e., random forest, confirming the alternative hypothesis (H_1) under RQ-1.1 stated in the Section 3.2. The inherent heavy data imbalance does not allow comparison with typical segmentation task performance. However, the results were comparable to the relevant work by Oehmcke et al. (2016), which reported an F1-score of 35.70% and an IoU of 21.70% in detecting small roads from Sentinel-2 imagery using a discriminant model. Overfitting is a particularly challenging task in improving the performance of DL models learning from such strongly imbalanced data. Accordingly, deeper and intuitive convolutional networks with less tendency to overfit, such as DeepUNET (Li et al., 2018), should be explored in future related studies to improve DL model learning capacity.

Moreover, the study also compared the use of Sentinel-2 imagery with that of Planet imagery in examining the relative potential of the Sentinel-2 imagery in the task. The results showed that Sentinel-2 images were more suitable than the Planet one in pixel-wise classification road pavement quality, reporting a higher F1-score (37.93% > 32.40%) and IoU (36.54% > 31.75%). This outcome rejects the null hypothesis (H_0) under RQ-2.1 and confirms the alternative, which stated that using Planet imagery would yield equal or worse performance in extracting RPQ compared to using Sentinel-2 imagery. The higher performance of using Sentinel-2 imagery can be attributed to the additional two bands in *SWIR1* and *SWIR2*, which the Planet imagery does not possess. These bands have previously been identified as indicators of road deterioration (Herold et al., 2004, 2008; Özdemir et al., 2020).

The proposed methodology of integrating spectral unmixing as a pre-processing step for Sentinel-2 images proved to have a degrading effect on the performance of the models. Abundance maps, i.e., the outputs of spectral unmixing, stacked together with the raw Sentinel-2 image, showed significantly lower performance compared to that of only using the raw bands. This results implies that the null hypothesis (H_0) under RQ-3.1, which stated that applying spectral unmixing as a pre-processing step would not improve the performance of the models in mapping RPQ is confirmed. Upon assessment of the intermediate products, i.e., the spectral features of endmembers and the abundance maps, the cause for the performance loss was identified to be the inconsistency in the band information structure of the abundance maps after endmember clustering. An abundance map corresponding to a certain endmember in one patch might end up representing another endmember in another patch. This inconsistency tends to confuse the ML algorithms and thus produce poor results. In this regard, this research recommends the exploration of more advanced unmixing techniques that can maintain consistency in structuring the endmembers globally. One such method can be the multiple endmember spectral mixture analysis (MESMA) method (Roberts et al., 1998) or its variants for a more automated procedure.

Furthermore, the study assessed several sources of uncertainty related to the applicability of the proposed methodology. In addressing these uncertainties, several avenues of model analysis paradigms were explored. Using Street View as an additional degree of observation, a slightly more robust validation approach was established. The results of this validation revealed that a significant portion of the discrepancy between the model predictions and reference data could be explained by timeline differences and spatial granularity. Comparing several spectral indices of the RPQ classes from the reference data and predictions indicated that the best performing model is, for the most part (i.e., since there were outliers), learning to recognize the spectral characteristics of each class. The model's capacity to detect defects was also assessed through cross-checking with the RPQ annotations. This result showed that some defects, such as rumble strips, which have a larger spatial coverage, are more frequently and correctly detected than others. These assessments helped to understand the limitations of the methodology. However, with these uncertainties in mind, it is recommended that the proposed approach be repeated on more reliably extracted and time-aligned reference data to realize the actual value of the methodology.

Towards addressing the uncertainty regarding severity classes, a three-class RPQ classification model was presented. This model showed significantly improved performance with an F1-score of 53.65% and an IoU of 46.03%. Alternative to this approach, a flexible modeling paradigm in the form of probabilistic threshold moving was also explored. Assisted with heuristics of the precision-recall curve and the probabilistic nature of ML model predictions, the study showed that predictions of the models could be molded to suit the utility desired. This approach is grounded on the trade-off between precision and recall, which translates to the trade-off between 'false alarm' rate and incompleteness of information. Models can be developed using this approach to be applicable to various users under the basis of a single architecture.

7.1. Reflection on the wicked problem of PHLs

Significant value can be obtained in extending the results above towards the social challenge considered in this study. The proposed methodology would enable regular road infrastructure monitoring at minimal cost, which is an essential attribute for resource strained regions challenged by PHLs because of poor road quality. Primarily, it can be used by road authorities at a higher level to make well-informed decisions on further detailed investigations and future rehabilitation plans based on granular observations of RPQ obtained from Sentinel-2 imagery. With frequent observations enabled by the Sentinel-2, road quality deterioration can be monitored conveniently to outline optimal rehabilitation plans. Truckers (transporters) can plot out their travel plans based on RPQ maps. Since it can ideally offer road quality information at IQL-4, it can link the road users with managing authorities, thereby serving as a check and balance in road infrastructure management. These contributions will collectively induce better awareness and management of road quality, which translates to reduced PHLs caused by transportation damages. In a more systemic and directed way, it is also possible to use the product of this methodology as an input to develop models of estimating PHL and vehicle maintenance costs, offering a seamless information infrastructure for food supply chain management and decision-making.

However, these applications are predicated on the decent performance of the proposed methodology. In that regard, it is necessary to address the uncertainties pointed out in this study. Notably, more work is required to isolate the road section using advanced road extraction algorithms. The RPQ segmentation problem is simpler to address once the non-road pixels are clipped out. Moreover, the generalizability of the method has to be scaled by training the models on data from various regions and conditions. The models can be applied in broader contexts and in various areas that can benefit from the approach with increased generalization capability.

APPENDIX

Annex 1 - Taxonomy and description of road surface distress

Table A 7.1 Taxonomy and description of road surface distresses taken from Paterson (1990)

Mode	Type	Brief description
Cracking	Crocodile	- Interconnected polygons of less than 300 mm diameter
	Longitudinal	- Line cracks longitudinal along the pavement
	Transverse	- Line cracks transverse across the pavement
	Irregular	- Unconnected cracks without a distinct pattern
	Map	- Interconnected polygons more than 300 mm in diameter
	Block	- Intersecting line cracks in a rectangular pattern at spacing greater than 1m
Disintegration	Raveling	- Loss of stone particles from surfacing
	Potholes	- Open cavity in surfacing (> 150 mm dia. and 50 mm depth)
	Edge break	- Loss of fragments at the edge of surfacing
Deformation	Rut	- Longitudinal depression in the wheel path
	Depression	- Bowl-shaped depression in surfacing
	Mound	- Localized rise in surfacing
	Ridge	- Longitudinal rise in surfacing
	Corrugation	- Transverse depression at close spacing
	Undulation	- Transverse depressions at long spacing (> 5 m)
	Roughness	- Irregularity of pavement surface in the wheel path

Annex 2 Data Specifications and pre-processing configuration

Table A 7.2 Sentinel-2 image tiles used and their date of ingestion

Sentinel Product	Date of ingestion	Tile ID
MSIL2C	1/7/2017	T30PXU
MSIL2C	1/7/2017	T30PYQ
MSIL2C	1/27/2017	T30NXN
MSIL2C	1/27/2017	T30NYN
MSIL2C	1/27/2017	T30PXS
MSIL2C	1/27/2017	T30PXT
MSIL2C	1/27/2017	T30PYR
MSIL2C	1/27/2017	T30PYS
MSIL2C	1/27/2017	T30PYT
MSIL2C	1/27/2017	T30PYU
MSIL2C	1/30/2017	T30PXU
MSIL2C	3/18/2017	T30PXQ
MSIL2C	3/18/2017	T30PXR
MSIL2C	4/17/2017	T30NYM
MSIL2C	12/23/2017	T30NXP
MSIL2C	12/15/2017	T30NZM

Table A 7.3 Sen2Cor atmospheric correction configuration

Atmospheric correction lookup table	Configuration	Remark
Aerosol type	AUTO	
Mid_latitude	AUTO	
Ozone content	0	To get the best approximation from metadata

Annex 3 Details of the SPICE unmixing algorithm

The SPICE method is a sparsity-promoting extension of the iterative constrained endmember extraction algorithm (ICE) (Berman et al., 2004). The ICE algorithm is a statistical approach that works by minimizing the error between the original pixel spectra and the algorithm estimates via the minimization of a residual sum of squares (RSS) term based on the convex geometry model (see Section 2.3.1) (Berman et al., 2004). Breman et al. (2004) added a sum of squared distances (SSD) term to the objective function, arguing that the solution of minimizing the RSS terms is not unique. Since the SSD term is related to the volume bounded by the endmembers (i.e., the simplex volume), adding this term to the objective function entails that the algorithm finds endmembers that provide a tight fit around the data (Zare, 2008). Breman et al. (2004) formulates this term as shown in Equation (7-1),

$$SSD = M\,(M-1)\,V \tag{7-1}$$

where M is the number of endmembers and V is the sum (over the bands) of the variances of the simplex vertices. However, instead of adding the SSD term, ICE uses V in the objective function to make it independent of the number of endmembers (M) (Berman et al., 2004). Moreover, a “regularization” parameter, μ, is included in the function to control the trade-off between the RSS and SSD terms. Accordingly, the ICE objective function is as shown in Equation (7-2),

$$RSS_{reg} = (1-\mu)\,\frac{RSS}{N} + \mu V \tag{7-2}$$

When it comes to the SPICE algorithm, the goal of sparsity promotion over the ICE method is to minimize the number of parameters generally applied by driving parameter values to zero (Zare, 2008). Typically, especially in neural network models, this is done by adding a weight decay term to the objective function (Williams, 1995). Similarly, the SPICE algorithm adds a sparsity-promoting term (SPT) of the form shown in Equation (7-3) to the ICE objective function.

$$SPT = \sum_{k=1}^{M} \gamma_k \sum_{i=1}^{N} p_{ik} \tag{7-3}$$

where N is the number of pixels in the image, p_{ik} is the proportion of endmember k in pixel i, and γ_k is set as shown in Equation (7-4),

$$\gamma_k = \frac{\Gamma}{\sum_{i=1}^{N} p_{ik}} \tag{7-4}$$

where Γ is a constant that controls the degree that the proportion values are driven to zero (i.e., sparsity). Since Γ is a constant throughout the minimization of the objective function, if the sum of a certain endmember’s proportion values (abundance) becomes small, then its weight, γ_k, becomes larger (i.e., inverse relationship). This weight increase encourages the minimization of the corresponding proportion values (abundances). Including the SPT into the ICE objective function, Zare and Gader (2007) formulates the objective function of SPICE as shown in Equation (7-5),

$$RSS_{reg}^{*} = (1-\mu)\frac{RSS}{N} + \mu V + SPT \tag{7-5}$$

The SPICE algorithm minimizes this function iteratively, checking, after each iteration, the maximum proportions for every endmember (Equation (7-6)). If this maximum proportion drops below a user-defined threshold, the endmember can be removed (pruned) from the set of endmembers (Zare & Gader, 2007).

$$MAXP_k = \max_i \{p_{ik}\} \tag{7-6}$$

Table A 7.4 Parameter configuration for SPICE unmixing

Parameter	**Recommended values**	**Used values**	**Remarks**
μ	0.001-0.1	0.01	To account for some degree of noise assumed to be in the data
Γ (gamma)	1-10	1	Empirically (trial & error) observing global consistency of endmembers
Initial no. of endmembers	20	20	As recommended
Objective change threshold	-	0.001	Based on desired numerical precision
Endmember pruning threshold	-	0.0005	To reduce excess endmember variability since it is applied on patches

Annex 4 Endmember clustering GMM selection

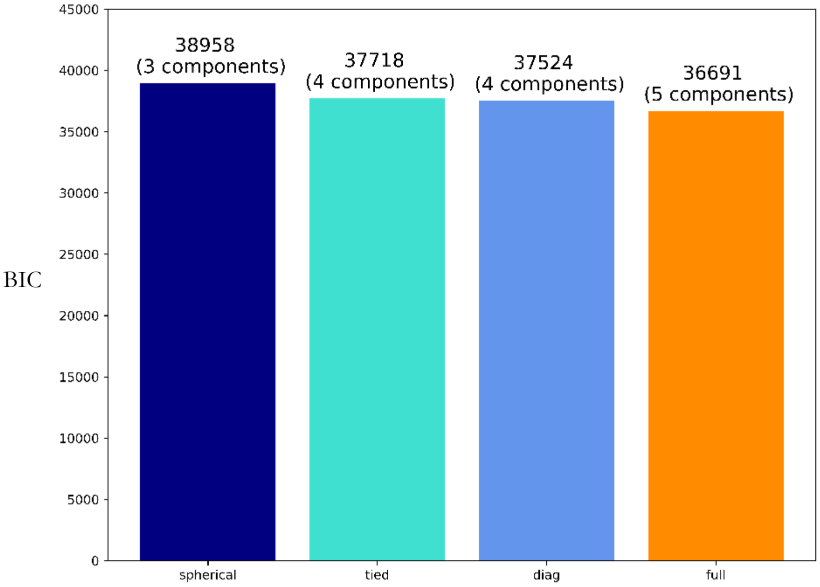


Figure A 7.1 GMM cluster model selection

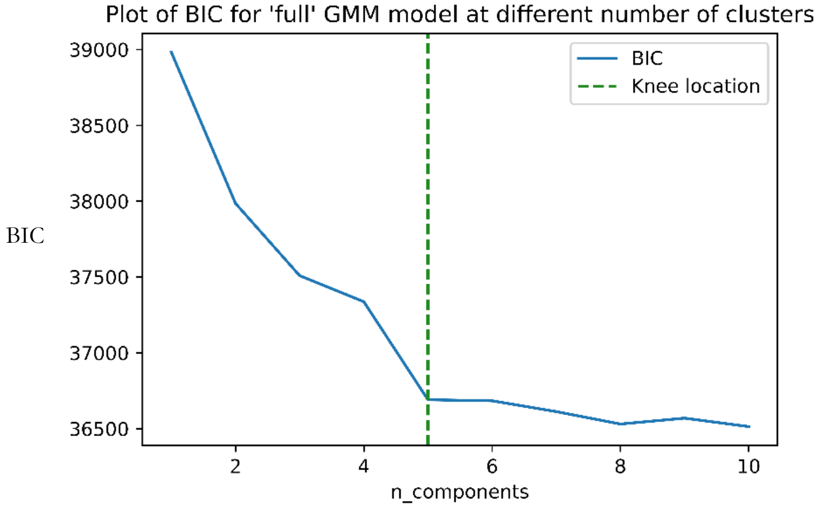


Figure A 7.2 Knee point location for 'full' covariance GMM model BIC plot

Annex 5 – U-Net model specifications

Table A 7.5 Architecture of the proposed U-Net model

U-Net Encoder			U-Net Decoder		
Layer	**Details**	**Size**	**Layer**	**Details**	**Size**
Input	Input Image	96x96x6	TranspConv1	kernel size: 3x3 stride: 2; filters: 512 BatchNorm LeakyReLu: 0.3 concat: Conv4_1	12x12x512
Conv1_1	kernel size: 3x3 filters: 64 BatchNorm LeakyReLu: 0.3	96x96x64	Conv6_1	kernel size: 3x3 filters: 512 BatchNorm LeakyReLu: 0.3	12x12x512
Conv1_2	kernel size: 3x3 filters: 64 BatchNorm LeakyReLu: 0.3	96x96x64	Conv6_2	kernel size: 3x3 filters: 512 BatchNorm LeakyReLu: 0.3	12x12x512
Pool1	2x2 max pool stride: 2	48x48x64	TranspConv2	kernel size: 3x3 stride: 2; filters: 256 BatchNorm LeakyReLu: 0.3 concat: Conv3_1	24x24x256
Conv2_1	kernel size: 3x3 filters: 64 BatchNorm LeakyReLu: 0.3	48x48x128	Conv7_1	kernel size: 3x3 filters: 256 BatchNorm LeakyReLu: 0.3	24x24x256
Conv2_2	kernel size: 3x3 filters: 64 BatchNorm LeakyReLu: 0.3	48x48x128	Conv7_2	kernel size: 3x3 filters: 256 BatchNorm LeakyReLu: 0.3	24x24x256
Pool2	2x2 max pool stride: 2	24x24x128	TranspConv2	kernel size: 3x3 stride: 2; filters: 128 BatchNorm LeakyReLu: 0.3 concat: Conv2_1	48x48x128
Conv3_1	kernel size: 3x3 filters: 64 BatchNorm LeakyReLu: 0.3	24x24x256	Conv8_1	kernel size: 3x3 filters: 128 BatchNorm LeakyReLu: 0.3	48x48x128
Conv3_2	kernel size: 3x3 filters: 64 BatchNorm LeakyReLu: 0.3	24x24x256	Conv8_2	kernel size: 3x3 filters: 128 BatchNorm LeakyReLu: 0.3	48x48x128
Pool3	2x2 max pool stride: 2	12x12x256	TranspConv3	kernel size: 3x3 stride: 2; filters: 64 BatchNorm	96x96x64

U-Net Encoder			**U-Net Decoder**		
Layer	**Details**	**Size**	**Layer**	**Details**	**Size**
				LeakyReLu: 0.3 concat: Conv1_1	
Conv4_1	kernel size: 3x3 filters: 64 BatchNorm LeakyReLu: 0.3	12x12x512	Conv9_1	kernel size: 3x3 filters: 64 BatchNorm LeakyReLu: 0.3	96x96x64
Conv4_2	kernel size: 3x3 filters: 64 BatchNorm LeakyReLu: 0.3	12x12x512	Conv9_2	kernel size: 3x3 filters: 64 BatchNorm LeakyReLu: 0.3	96x96x64
Pool4	2x2 max pool stride: 2	6x6x512	Conv10	kernel size: 1x1 filters: 4 SoftMax	96x96x4
Conv5_1	kernel size: 3x3 filters: 64 BatchNorm LeakyReLu: 0.3	6x6x1024			
Conv5_2	kernel size: 3x3 filters: 64 BatchNorm LeakyReLu: 0.3	6x6x1024			

Annex 6 Hyperparameter configurations

Table A 7.6 Selected hyperparameters for the random forest model

Hyperparameter	**Selected configuration**
Initial number of estimators (trees)	150
Feature selection criterion	Gini
Maximum depth	8
Maximum features	Sqrt of number features
bootstrap	No
Mini-batch size	200
Estimators added each batch	50

Table A 7.7 Selected hyperparameter for the U-Net model

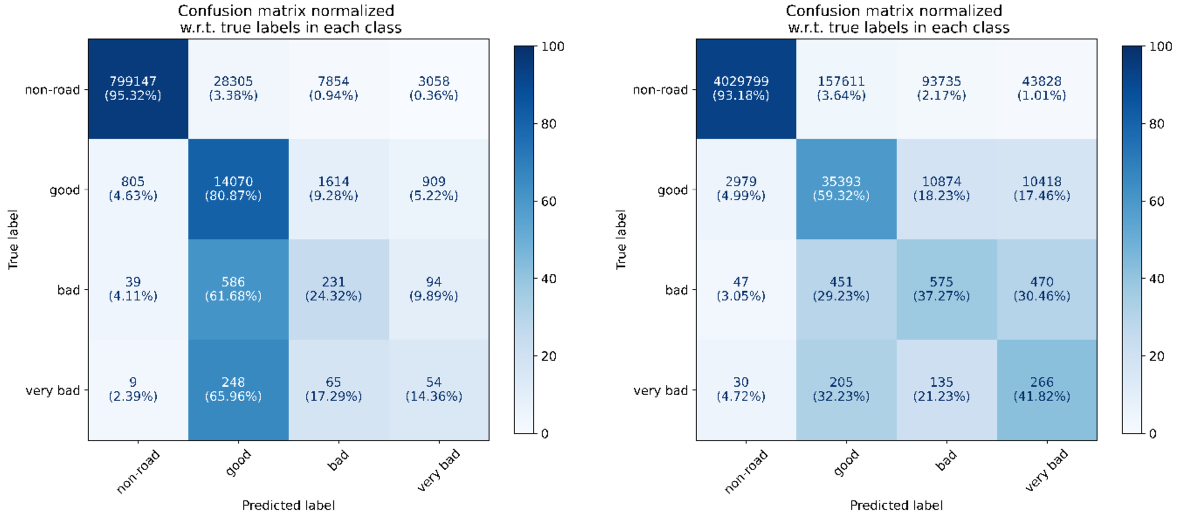
Hyperparameter	**Selected configuration**
Initial learning rate	10^{-5}
Batch size*	5 (2)
Learning rate decay step*	20 (10) epochs
Learning rate decay rate	0.5
L2 rate	0.1
L1 rate	0.0001
Focal loss alpha	0.5
Focal loss gamma	2

* The values in brackets are those applied for the Planet dataset

Annex 7 – Performance evaluation results

Table A 7.8 Class-wise performance results of all four class experiments

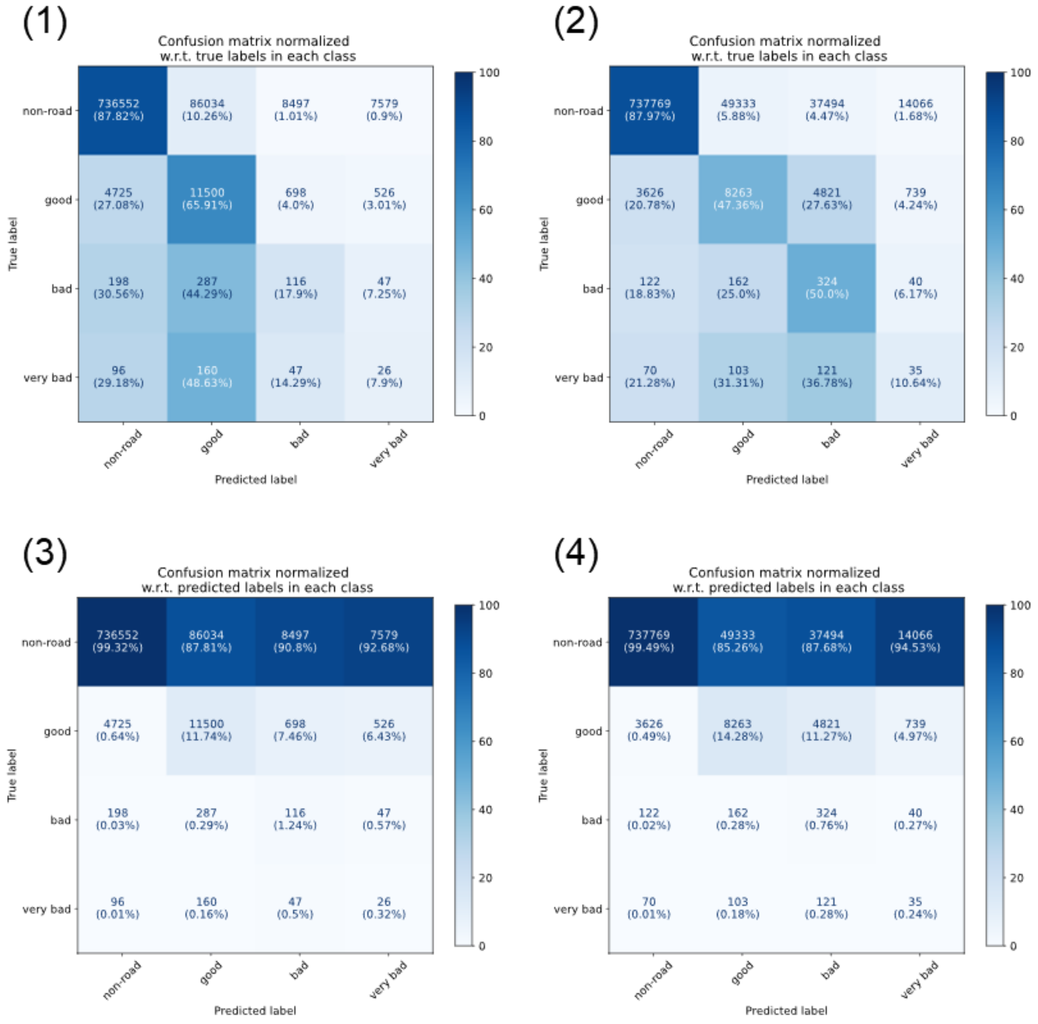
		Experiment							
Metrics	Class	RF_S	RF_SU	RF_SC	RF_P	DL_S	DL_SU	DL_SC	DL_P
P	Non-road	99.89%	99.32%	99.90%	99.92%	99.97%	99.49%	99.83%	99.92%
	Good	32.56%	11.74%	26.60%	18.28%	38.73%	14.28%	14.56%	37.50%
	Bad	2.37%	1.24%	1.18%	0.55%	2.63%	0.76%	0.90%	1.34%
	Very bad	1.31%	0.32%	0.84%	0.48%	1.18%	0.24%	0.41%	0.86%
R	Non-road	95.32%	87.82%	93.84%	93.18%	95.28%	87.97%	86.09%	97.31%
	Good	80.87%	65.91%	83.41%	59.32%	59.35%	47.36%	47.24%	51.11%
	Bad	24.32%	17.90%	12.76%	37.27%	56.63%	50.00%	25.13%	52.62%
	Very bad	14.36%	7.90%	14.22%	41.82%	34.04%	10.64%	53.05%	44.50%
F	Non-road	97.55%	93.22%	96.78%	96.43%	97.57%	93.37%	92.45%	98.60%
	Good	46.43%	19.93%	40.34%	27.94%	46.87%	21.94%	22.26%	43.26%
	Bad	4.31%	2.32%	2.17%	1.08%	5.02%	1.49%	1.74%	2.61%
	Very bad	2.40%	0.61%	1.59%	0.96%	2.28%	0.46%	0.82%	1.69%



	non-road	good	bad	very bad
non-road	799147 (95.32%)	28305 (3.38%)	7854 (0.94%)	3058 (0.36%)
good	805 (4.63%)	14070 (80.87%)	1614 (9.28%)	909 (5.22%)
bad	39 (4.11%)	586 (61.68%)	231 (24.32%)	94 (9.89%)
very bad	9 (2.39%)	248 (65.96%)	65 (17.29%)	54 (14.36%)

	non-road	good	bad	very bad
non-road	4029799 (93.18%)	157611 (3.64%)	93735 (2.17%)	43828 (1.01%)
good	2979 (4.99%)	35393 (59.32%)	10874 (18.23%)	10418 (17.46%)
bad	47 (3.05%)	451 (29.23%)	575 (37.27%)	470 (30.46%)
very bad	30 (4.72%)	205 (32.23%)	135 (21.23%)	266 (41.82%)

Figure A 7.3 True normalized confusion matrix of experiments RF_S (left) and RF_P (right)



(1) Confusion matrix normalized w.r.t. true labels in each class

True label \ Predicted label	non-road	good	bad	very bad
non-road	736552 (87.82%)	86034 (10.26%)	8497 (1.01%)	7579 (0.9%)
good	4725 (27.08%)	11500 (65.91%)	698 (4.0%)	526 (3.01%)
bad	198 (30.56%)	287 (44.29%)	116 (17.9%)	47 (7.25%)
very bad	96 (29.18%)	160 (48.63%)	47 (14.29%)	26 (7.9%)

(2) Confusion matrix normalized w.r.t. true labels in each class

True label \ Predicted label	non-road	good	bad	very bad
non-road	737769 (87.97%)	49333 (5.88%)	37494 (4.47%)	14066 (1.68%)
good	3626 (20.78%)	8263 (47.36%)	4821 (27.63%)	739 (4.24%)
bad	122 (18.83%)	162 (25.0%)	324 (50.0%)	40 (6.17%)
very bad	70 (21.28%)	103 (31.31%)	121 (36.78%)	35 (10.64%)

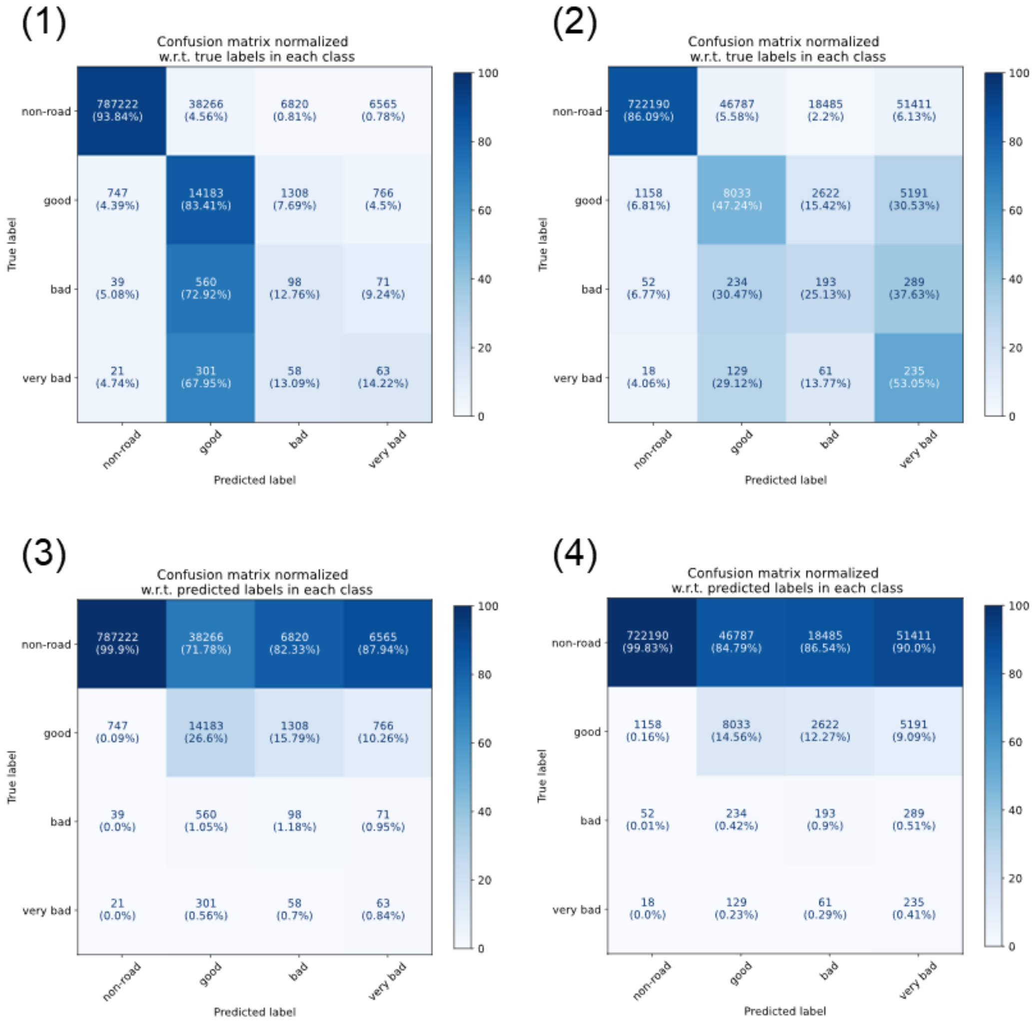
(3) Confusion matrix normalized w.r.t. predicted labels in each class

True label \ Predicted label	non-road	good	bad	very bad
non-road	736552 (99.32%)	86034 (87.81%)	8497 (90.8%)	7579 (92.68%)
good	4725 (0.64%)	11500 (11.74%)	698 (7.46%)	526 (6.43%)
bad	198 (0.03%)	287 (0.29%)	116 (1.24%)	47 (0.57%)
very bad	96 (0.01%)	160 (0.16%)	47 (0.5%)	26 (0.32%)

(4) Confusion matrix normalized w.r.t. predicted labels in each class

True label \ Predicted label	non-road	good	bad	very bad
non-road	737769 (99.49%)	49333 (85.26%)	37494 (87.68%)	14066 (94.53%)
good	3626 (0.49%)	8263 (14.28%)	4821 (11.27%)	739 (4.97%)
bad	122 (0.02%)	162 (0.28%)	324 (0.76%)	40 (0.27%)
very bad	70 (0.01%)	103 (0.18%)	121 (0.28%)	35 (0.24%)

Figure A 7.4 Confusion matrices of experiments RF_SU (1 & 3) and DL_SU (2 & 4)



(1)

	non-road	good	bad	very bad
non-road	787222 (93.84%)	38266 (4.56%)	6820 (0.81%)	6565 (0.78%)
good	747 (4.39%)	14183 (83.41%)	1308 (7.69%)	766 (4.5%)
bad	39 (5.08%)	560 (72.92%)	98 (12.76%)	71 (9.24%)
very bad	21 (4.74%)	301 (67.95%)	58 (13.09%)	63 (14.22%)

(2)

	non-road	good	bad	very bad
non-road	722190 (86.09%)	46787 (5.58%)	18485 (2.2%)	51411 (6.13%)
good	1158 (6.81%)	8033 (47.24%)	2622 (15.42%)	5191 (30.53%)
bad	52 (6.77%)	234 (30.47%)	193 (25.13%)	289 (37.63%)
very bad	18 (4.06%)	129 (29.12%)	61 (13.77%)	235 (53.05%)

(3)

	non-road	good	bad	very bad
non-road	787222 (99.9%)	38266 (71.78%)	6820 (82.33%)	6565 (87.94%)
good	747 (0.09%)	14183 (26.6%)	1308 (15.79%)	766 (10.26%)
bad	39 (0.0%)	560 (1.05%)	98 (1.18%)	71 (0.95%)
very bad	21 (0.0%)	301 (0.56%)	58 (0.7%)	63 (0.84%)

(4)

	non-road	good	bad	very bad
non-road	722190 (99.83%)	46787 (84.79%)	18485 (86.54%)	51411 (90.0%)
good	1158 (0.16%)	8033 (14.56%)	2622 (12.27%)	5191 (9.09%)
bad	52 (0.01%)	234 (0.42%)	193 (0.9%)	289 (0.51%)
very bad	18 (0.0%)	129 (0.23%)	61 (0.29%)	235 (0.41%)

Figure A 7.5 Confusion matrices of experiments RF_SC (1 & 3) and DL_SC (2 & 4)

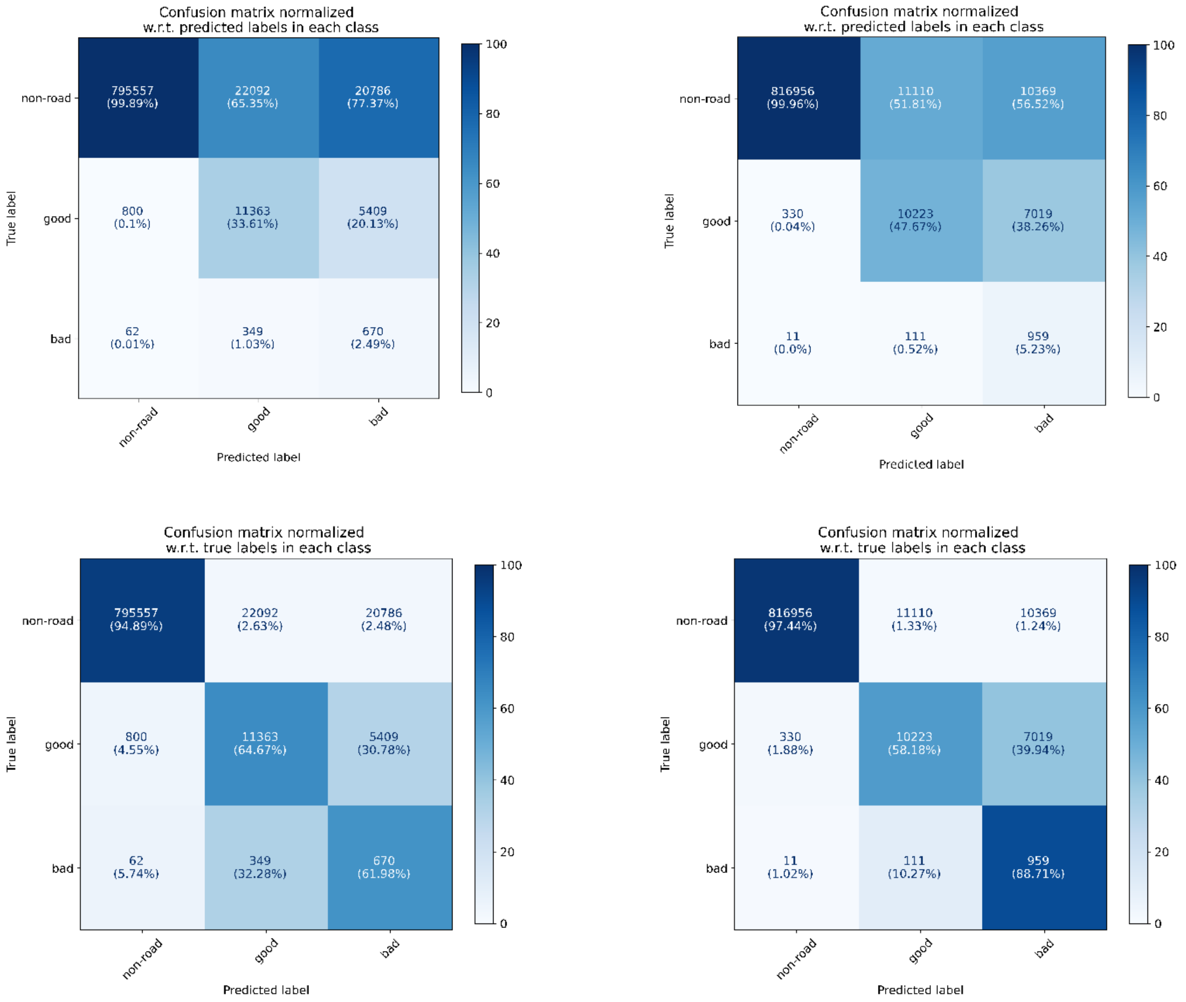


Figure A 7.6 Confusion matrices of experiments RF_S_3C (1 & 3) and DL_S_3C (2 & 4)

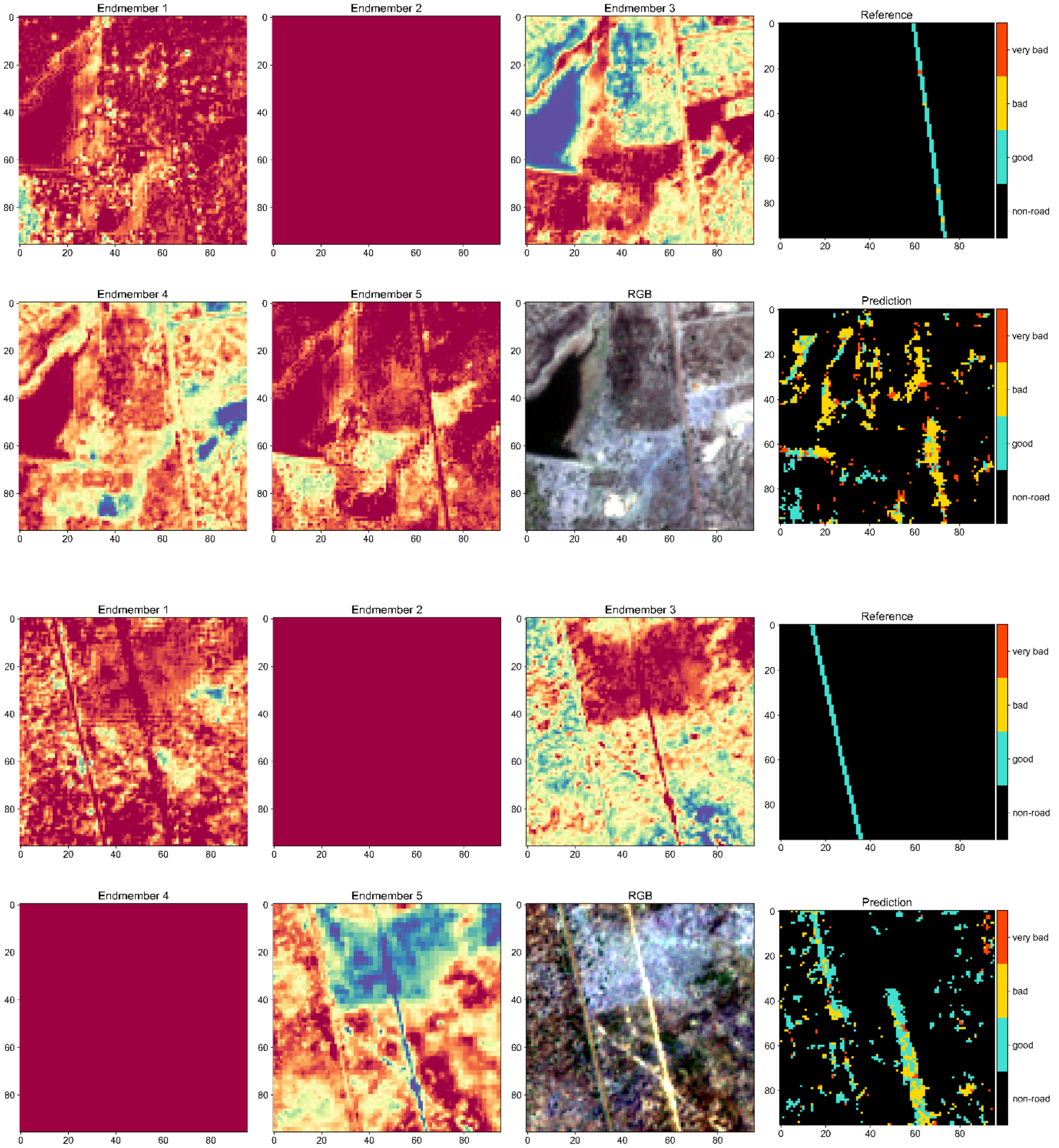


Figure A 7.7 Endmember abundance maps and model predictions of DL_SU compared to the reference labels

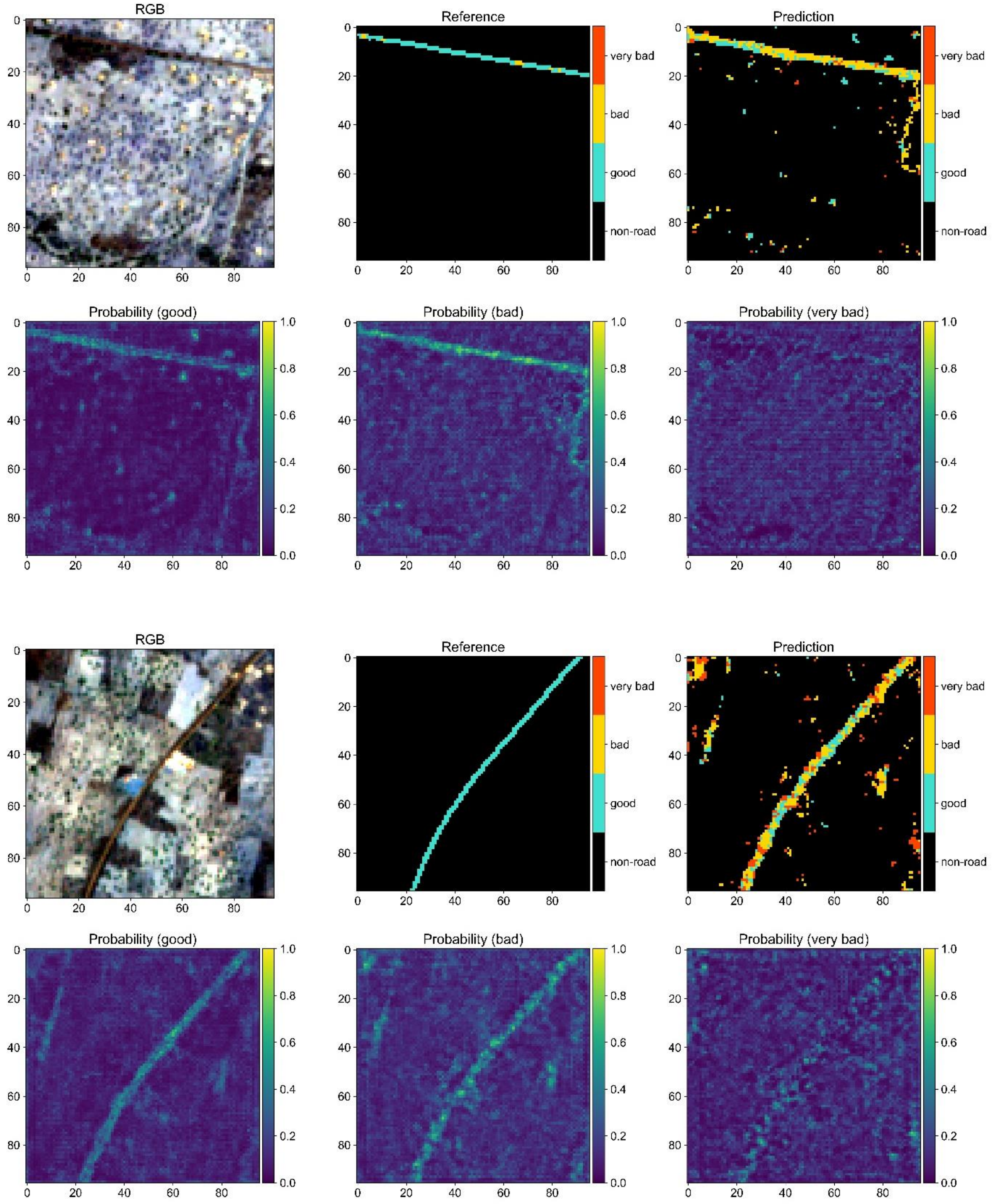


Figure A 7.8 Plots showing predictions with their class-wise probabilities

LIST OF REFERENCES

Abdellatif, M., Peel, H., Cohn, A. G., & Fuentes, R. (2019). Hyperspectral imaging for autonomous inspection of road pavement defects. *Proceedings of the 36th International Symposium on Automation and Robotics in Construction, ISARC 2019*, 384–392. https://doi.org/10.22260/isarc2019/0052

Affognon, H., Mutungi, C., Sanginga, P., & Borgemeister, C. (2015). Unpacking postharvest losses in sub-Saharan Africa: A Meta-Analysis. *World Development*, *66*, 49–68. https://doi.org/10.1016/j.worlddev.2014.08.002

Ager, T. (2013). An Introduction to Synthetic Aperture Radar Imaging. *Oceanography*, *26*(2). https://doi.org/10.5670/oceanog.2013.28

Ait Kerroum, M., Hammouch, A., & Aboutajdine, D. (2010). Textural feature selection by joint mutual information based on Gaussian mixture model for multispectral image classification. *Pattern Recognition Letters*, *31*(10), 1168–1174. https://doi.org/10.1016/j.patrec.2009.11.010

Akaike, H. (1972). Information theory and an extension of the maximum likelihood principle. *Proc. 2nd Int. Symp. Information Theory, Supp. to Problems of Control and Information Theory*, 267–281.

Alam, F. I., Zhou, J., Tong, L., Liew, A. W.-C., & Gao, Y. (2017). Combining Unmixing and Deep Feature Learning for Hyperspectral Image Classification. *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, *2017-Decem*, 1–8. https://doi.org/10.1109/DICTA.2017.8227419

Andreou, C., Karathanassi, V., & Kolokoussis, P. (2011). Investigation of hyperspectral remote sensing for mapping asphalt road conditions. *International Journal of Remote Sensing*, *32*(21), 6315–6333. https://doi.org/10.1080/01431161.2010.508799

Ari, C., & Aksoy, S. (2014). Detection of Compound Structures Using a Gaussian Mixture Model With Spectral and Spatial Constraints. *IEEE Transactions on Geoscience and Remote Sensing*, *52*(10), 6627–6638. https://doi.org/10.1109/TGRS.2014.2299540

Bellù, L. G. (2017). *Food losses and waste: issues and policy option*. www.fao.org/publications

Bennett, C. R., de Solminihac, H., & Chamorro, A. (2006). *Data Collection Technologies for Road Management*. https://openknowledge.worldbank.org/handle/10986/11776

Bennett, C. R., & Paterson, W. D. O. (2000). A Guide to Calibration and Adaptation. HDM-4. Volume 5. In *Highway Developmenet & Management Series*. PIARC.

Bergström, J., & Dekker, S. W. A. (2014). Bridging the macro and the micro by considering the meso: Reflections on the fractal nature of resilience. *Ecology and Society*, *19*(4), 22. https://doi.org/10.5751/ES-06956-190422

Berman, M., Kiiveri, H., Lagerstrom, R., Ernst, A., Dunne, R., & Huntington, J. F. (2004). ICE: a statistical approach to identifying endmembers in hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing*, *42*(10), 2085–2095. https://doi.org/10.1109/TGRS.2004.835299

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, *10*(2–3), 191–203. https://doi.org/10.1016/0098-3004(84)90020-7

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning* (1st ed.). Springer-Verlag New York.

Bollen, A. F., Riden, C. P., & Opara, L. U. (2006). Traceability in postharvest quality management. *International Journal of Postharvest Technology and Innovation*, *1*(1), 93. https://doi.org/10.1504/IJPTI.2006.009186

Borsoi, R. A., Imbiriba, T., Bermudez, J. C. M., Richard, C., Chanussot, J., Drumetz, L., Tourneret, J. Y., Zare, A., & Jutten, C. (2020). Spectral variability in hyperspectral data unmixing: A comprehensive review. In *arXiv*. arXiv. https://github.com/ricardoborsoi/

Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. In *ACM Computing Surveys* (Vol. 49, Issue 2, pp. 1–50). https://doi.org/10.1145/2907070

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brewer, E., Lin, J., Kemper, P., Hennin, J., & Runfola, D. (2021). Predicting road quality using high resolution satellite imagery: A transfer learning approach. *PLOS ONE*, *16*(7), e0253370. https://doi.org/10.1371/journal.pone.0253370

Bridgelall, R., Rafert, J. B., & Tolliver, D. (2015). Hyperspectral imaging utility for transportation systems. In J. P. Lynch (Ed.), *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2015* (Vol. 9435, p. 943522). SPIE. https://doi.org/10.1117/12.2083957

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The Balanced Accuracy and Its Posterior Distribution. *2010 20th International Conference on Pattern Recognition*, 3121–3124. https://doi.org/10.1109/ICPR.2010.764

Brownlee, J. (2020). *A Gentle Introduction to Imbalanced Classification*. Machine Learning Mastery. https://machinelearningmastery.com/what-is-imbalanced-classification/#:~:text=Imbalanced classification refers to a,is instead biased or skewed.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B., & Varoquaux, G. (2013). *API design for machine learning software: experiences from the scikit-learn project*. https://arxiv.org/abs/1309.0238v1

Cadamuro, G., Muhebwa, A., & Taneja, J. (2018). *Assigning a Grade: Accurate Measurement of Road Quality Using Satellite Imagery*.

Cadamuro, G., Muhebwa, A., & Taneja, J. (2019). Street Smarts: Measuring Intercity Road Quality Using Deep Learning on Satellite Imagery. *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS 19)*, 145–154. https://doi.org/10.1145/3314344.3332493

Çelik, T. (2011). Bayesian change detection based on spatial sampling and Gaussian mixture model. *Pattern Recognition Letters*, *32*(12), 1635–1642. https://doi.org/10.1016/j.patrec.2011.05.008

Chaboud, G., & Daviron, B. (2017). Food losses and waste: Navigating the inconsistencies. *Global Food Security*, *12*, 1–7. https://doi.org/10.1016/j.gfs.2016.11.004

Chambon, S., & Moliard, J.-M. (2011). Automatic Road Pavement Assessment with Image Processing: Review and Comparison. *International Journal of Geophysics*, *2011*, 1–20. https://doi.org/10.1155/2011/989354

Chen, J., Zhu, W., Tian, Y. Q., Yu, Q., Zheng, Y., & Huang, L. (2017). Remote estimation of colored dissolved organic matter and chlorophyll-a in Lake Huron using Sentinel-2 measurements. *Journal of Applied Remote Sensing*, *11*(03), 1. https://doi.org/10.1117/1.jrs.11.036007

Chen, S.-L., Lin, C.-H., Tang, C.-W., Chu, L.-P., & Cheng, C.-K. (2020). Research on the International Roughness Index Threshold of Road Rehabilitation in Metropolitan Areas: A Case Study in Taipei City. *Sustainability*, *12*(24), 10536. https://doi.org/10.3390/su122410536

Chen, X.-L., Zhao, H.-M., Li, P.-X., & Yin, Z.-Y. (2006). Remote sensing image-based analysis of the relationship between urban heat island and land use/cover changes. *Remote Sensing of Environment*, *104*(2), 133–146. https://doi.org/10.1016/j.rse.2005.11.016

Cheng, G., Wang, Y., Xu, S., Wang, H., Xiang, S., & Pan, C. (2017). Automatic Road Detection and Centerline Extraction via Cascaded End-to-End Convolutional Neural Network. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(6), 3322–3337. https://doi.org/10.1109/TGRS.2017.2669341

Chonhenchob, V., Sittipod, S., Swasdee, D., Rachtanapun, P., Singh, S., & Singh, J. (2009). Effect of Truck Vibration during Transport on Damage to Fresh Produce Shipments in Thailand. *Journal of Applied Packaging Research*, *3*(1), 27.

Cohen, G., Hilario, M., Sax, H., Hugonnet, S., & Geissbuhler, A. (2006). Learning from imbalanced data in surveillance of nosocomial infection. *Artificial Intelligence in Medicine*, *37*(1), 7–18. https://doi.org/10.1016/j.artmed.2005.03.002

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Eriksson, J., Girod, L., Hull, B., Newton, R., Madden, S., & Balakrishnan, H. (2008). The Pothole Patrol: Using a mobile sensor network for road surface monitoring. *MobiSys'08 - Proceedings of the 6th International Conference on Mobile Systems, Applications, and Services*, 29–39. https://doi.org/10.1145/1378600.1378605

European Space Agency. (2015). Sentinel-2 User Handbook. In *ESA Standard Document* (Issue 1).

Fadiji, T., Coetzee, C., Chen, L., Chukwu, O., & Opara, U. L. (2016). Susceptibility of apples to bruising inside ventilated corrugated paperboard packages during simulated transport damage. *Postharvest Biology and Technology*, *118*, 111–119. https://doi.org/10.1016/j.postharvbio.2016.04.001

Fagrhi, A., & Ozden, A. (2015). *Satellite Assessment and Monitoring for Pavement Management* (Issue November). https://cait.rutgers.edu/wp-content/uploads/2018/05/cait-utc-nc4-final_0.pdf

FAO. (2011a). *Continental Programme on Post-Harvest Losses (PHL) Reduction - Reduction Rapid Country Needs Assessment (Ghana).*

FAO. (2011b). *Global food losses and food waste - Extent, causes and prevention.*

FAO. (2014). *Definitional Framework of Food Loss - Save Food: Global Initiative on Food Loss and Waste Reduction.*

FAO. (2017). The Future of Food and Agriculture. In *Food and Agriculture Organization of the United Nations* (Issue November). http://www.fao.org/3/i6583e/i6583e.pdf

FAO. (2018). *World Food Day 2018. Our Actions are our Future.*

FAO, IFAD, UNICEF, WFP, & WHO. (2020). *The State of Food Security and Nutrition in the World 2020. Transforming food systems for affordable healthy diets* (Vol. 7, Issue 7). FAO. https://doi.org/10.4060/ca9692en

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from Imbalanced Data Sets*. Springer International Publishing. https://doi.org/10.1007/978-3-319-98074-4

Fernando, I., Fei, J., Stanley, R., Enshaei, H., & Eyles, A. (2019). Quality deterioration of bananas in the post-harvest supply chain-an empirical study Quality deterioration of bananas. *Modern Supply Chain Research and Applications*, *1*(2), 2631–3871. https://doi.org/10.1108/MSCRA-05-2019-0012

Gao, B. C. (1996). NDWI - A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, *58*(3), 257–266. https://doi.org/10.1016/S0034-4257(96)00067-3

Gao, L., Song, W., Dai, J., & Chen, Y. (2019). Road Extraction from High-Resolution Remote Sensing Imagery Using Refined Deep Residual Convolutional Neural Network. *Remote Sensing*, *11*(5), 552. https://doi.org/10.3390/rs11050552

Georgiadou, Y., & Reckien, D. (2018). Geo-Information Tools, Governance, and Wicked Policy Problems. *ISPRS International Journal of Geo-Information*, *7*(1), 21. https://doi.org/10.3390/ijgi7010021

Godfray, H. C. J., Beddington, J. R., Crute, I. R., Haddad, L., Lawrence, D., Muir, J. F., Pretty, J., Robinson, S., Thomas, S. M., & Toulmin, C. (2010). Food security: The challenge of feeding 9 billion people. In *Science* (Vol. 327, Issue 5967, pp. 812–818). https://doi.org/10.1126/science.1185383

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. The MIT Press.

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for Multi-Class Classification: an Overview*. 1–17. http://arxiv.org/abs/2008.05756

Grolleaud, M. (2002). *Post-harvest Losses: Discovering the Full Story. Overview of the Phenomenon of Losses during the Post-harvest System*. Agricultural Support Systems Div. eng. http://www.fao.org/3/AC301E/AC301E00.htm

Gupta, A., Tatbul, N., Marcus, R., Zhou, S., Lee, I., & Gottschlich, J. (2020). *Class-Weighted Evaluation Metrics for Imbalanced Data Classification*. http://arxiv.org/abs/2010.05995

Guyon, I., Bennett, K., Cawley, G., Escalante, H. J., Escalera, S., Tin Kam Ho, Macia, N., Ray, B., Saeed, M., Statnikov, A., & Viegas, E. (2015). Design of the 2015 ChaLearn AutoML challenge. *2015 International Joint Conference on Neural Networks (IJCNN)*, *2015-Septe*, 1–8. https://doi.org/10.1109/IJCNN.2015.7280767

Hamwood, J., Alonso-Caneiro, D., Read, S. A., Vincent, S. J., & Collins, M. J. (2018). Effect of patch size and network architecture on a convolutional neural network approach for automatic segmentation of OCT retinal layers. *Biomedical Optics Express*, *9*(7), 3049. https://doi.org/10.1364/boe.9.003049

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Hastuti, E. Y. (2008). *The influence of agribusiness system applied to vegetables farmers income improvement in Boyolali regency*. Diponegoro University.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Herold, M., & Roberts, D. A. (2004). Mapping asphalt road condition with hyperspectral remote sensing. *Aviris 2004*, *August 2016*, 1–15.

Herold, M., Roberts, D., Noronha, V., & Smadi, O. (2008). Imaging spectrometry and asphalt road surveys. *Transportation Research Part C: Emerging Technologies*, *16*(2), 153–166. https://doi.org/10.1016/j.trc.2007.07.001

Herold, M., Roberts, D., Smadi, O., & Noronha, V. (2004). Road condition mapping with hyperspectral remote sensing. *Aviris 2004*, *August 2016*, 1–15. http://www.spyplanes.com/

HLPE. (2014). *Food losses and waste in the context of sustainable food systems: A report by The High Level Panel of Experts on Food Security and Nutrition of the Committee on World Food Security*. www.fao.org/cfs/cfs-hlpe

Hoppe, R. (2010). *The Governance of Problems: Puzzling, Powering and Participation*. Policy Press.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *32nd International Conference on Machine Learning, ICML 2015*, *1*, 448–456. http://arxiv.org/abs/1502.03167

Jarimopas, B., Singh, S. P., & Saengnil, W. (2005). Measurement and analysis of truck transport vibration levels and damage to packaged tangerines during transit. *Packaging Technology and Science*, *18*(4), 179–188. https://doi.org/10.1002/pts.687

Jarvis, A., Reuter, H. I., Nelson, A., & Guevara, E. (2008). *Hole-filled seamless SRTM data V4*. International Centre for Tropical Agriculture (CIAT). http://srtm.csi.cgiar.org

Jiang, L., Xie, Y., & Ren, T. (2020). A deep neural networks approach for pixel-level runway pavement crack segmentation using drone-captured images. *Transportation Research Board 99th Annual Meeting*, *April*, 1–13. https://arxiv.org/abs/2001.03257v1

Karimzadeh, S., & Matsuoka, M. (2021). Development of nationwide road quality map: Remote sensing meets field sensing. *Sensors*, *21*(6), 2251. https://doi.org/10.3390/s21062251

Kearney, J. (2010). Food consumption trends and drivers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1554), 2793–2807. https://doi.org/10.1098/rstb.2010.0149

Keshava, N., & Mustard, J. F. (2002). Spectral unmixing. *IEEE Signal Processing Magazine*, *19*(1), 44–57. https://doi.org/10.1109/79.974727

Kingma, D. P., & Ba, J. L. (2015, December 22). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. http://arxiv.org/abs/1412.6980

Kitinoja, L., & AlHassan, H. Y. (2012). Identification of Appropriate Postharvest Technologies for Small Scale Horticultural Farmers and Marketers in Sub-Saharan Africa and South Asia – Part 1. Postharvest Losses and Quality Assessments. *Acta Horticulturae*, *934*(934), 31–40. https://doi.org/10.17660/ActaHortic.2012.934.1

Kojo Arah, I., Kumah, E. K., Anku, E. K., & Amaglo, H. (2015). *An Overview of Post-Harvest Losses in Tomato Production in Africa: Causes and Possible Prevention Strategies*. *5*(16). www.iiste.org

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Lars Forslöf, & Hans Jones. (2015). Roadroid: Continuous Road Condition Monitoring with Smart Phones. *Journal of Civil Engineering and Architecture*, *9*(4), 485–496. https://doi.org/10.17265/1934-7359/2015.04.012

Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, *7*(1), 3–10. https://doi.org/10.1016/J.GSF.2015.07.003

Laubis, K. (2017). Crowd-Based Road Surface Monitoring and its Implications on Road Users and Road Authorities. In *Dissertation*. https://publikationen.bibliothek.kit.edu/1000080758

Laubis, K., Konstantinov, M., Simko, V., Gröschel, A., & Weinhardt, C. (2019). Enabling crowdsensing-based road condition monitoring service by intermediary. *Electronic Markets*, *29*(1), 125–140. https://doi.org/10.1007/s12525-018-0292-7

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Li, R., Liu, W., Yang, L., Sun, S., Hu, W., Zhang, F., & Li, W. (2018). DeepUNet: A Deep Fully Convolutional Network for Pixel-Level Sea-Land Segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *11*(11), 3954–3962. https://doi.org/10.1109/JSTARS.2018.2833382

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 318–327. https://arxiv.org/abs/1708.02002v2

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489

Londoño-Kent, P. (2009). *Freight Transport for Development Toolkit: Road Freight*. https://documents1.worldbank.org/curated/ar/649241468155983939/pdf/579680WP0road0Box353787B01PUBLIC1.pdf

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(4), 640–651. https://doi.org/10.1109/TPAMI.2016.2572683

Lu, L., Shin, Y., Su, Y., & Karniadakis, G. E. (2019). Dying ReLU and Initialization: Theory and Numerical Examples. *Communications in Computational Physics*, *28*(5), 1671–1706. https://doi.org/10.4208/cicp.oa-2020-0165

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *152*(April), 166–177. https://doi.org/10.1016/j.isprsjprs.2019.04.015

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. *In ICML Workshop on Deep Learning for Audio, Speech and Language Processing*.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *1*, 281–296.

Main-Knorn, M., Pflug, B., Louis, J., Debaecker, V., Müller-Wilm, U., & Gascon, F. (2017). Sen2Cor for Sentinel-2. In L. Bruzzone, F. Bovolo, & J. A. Benediktsson (Eds.), *Image and Signal Processing for Remote Sensing XXIII* (p. 3). SPIE. https://doi.org/10.1117/12.2278218

Maugis, C., Celeux, G., & Martin-Magniette, M.-L. (2009). Variable selection in model-based clustering: A general variable role modeling. *Computational Statistics & Data Analysis*, *53*(11), 3872–3882. https://doi.org/10.1016/j.csda.2009.04.013

Mettas, C., Themistocleous, K., Neocleous, K., Christofe, A., Pilakoutas, K., & Hadjimitsis, D. (2015). Monitoring asphalt pavement damages using remote sensing techniques. *Third International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2015)*, *9535*, 95350S. https://doi.org/10.1117/12.2195702

Meyer, F. J., Ajadi, O. A., & Hoppe, E. J. (2020). Studying the Applicability of X-Band SAR Data to the Network-Scale Mapping of Pavement Roughness on US Roads. *Remote Sensing 2020, Vol. 12, Page 1507*, *12*(9), 1507. https://doi.org/10.3390/RS12091507

Minarelli, F., Galioto, F., Raggi, M., & Viaggi, D. (2002). Asymmetric information along the food supply chain : a review of the literature. *Canadian Journal of Agricultural Economics*, *Volume 50*(Issue 1), 35–50.

Mnih, V., & Hinton, G. E. (2010). Learning to Detect Roads in High-Resolution Aerial Images. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 6316 LNCS* (Issue PART 6, pp. 210–223). https://doi.org/10.1007/978-3-642-15567-3_16

Mohan, P., Padmanabhan, V. N., & Ramjee, R. (2008). *Nericell: Rich Monitoring of Road and Traffic Conditions using Mobile Smartphones*.

Mubaraki, M. (2016). Highway subsurface assessment using pavement surface distress and roughness data. *International Journal of Pavement Research and Technology*, *9*(5), 393–402. https://doi.org/10.1016/j.ijprt.2016.10.001

Nelson, A., de Sherbinin, A., & Pozzi, F. (2006). Towards development of a high quality public domain global roads database. *Data Science Journal*, *5*, 223–265. https://doi.org/10.2481/DSJ.5.223

Oehmcke, S., Thrysøe, C., Borgstad, A., Antonio, M., Salles, V., Brandt, M., & Gieseke, F. (2016). *Detecting Hardly Visible Roads in Low-Resolution Satellite Time Series Data*.

Oehmcke, S., Thrysoe, C., Borgstad, A., Salles, M. A. V., Brandt, M., & Gieseke, F. (2019). Detecting Hardly Visible Roads in Low-Resolution Satellite Time Series Data. *2019 IEEE International Conference on Big Data (Big Data)*, 2403–2412. https://doi.org/10.1109/BigData47090.2019.9006251

Opara, U. L., & Pathare, P. B. (2014). Bruise damage measurement and analysis of fresh horticultural produce—A review. *Postharvest Biology and Technology*, *91*, 9–24. https://doi.org/10.1016/j.postharvbio.2013.12.009

O'Shea, K., & Nash, R. (2015). *An Introduction to Convolutional Neural Networks*. https://arxiv.org/abs/1511.08458v2

Oshri, B., Hu, A., Adelson, P., Chen, X., Dupas, P., Weinstein, J., Burke, M., Lobell, D., & Ermon, S. (2018). Infrastructure Quality Assessment in Africa using Satellite Imagery and Deep Learning. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, *1*(1), 616–625. https://doi.org/10.1145/3219819.3219924

Owusu, K., & Waylen, P. R. (2013). The changing rainy season climatology of mid-Ghana. *Theoretical and Applied Climatology*, *112*(3–4), 419–430. https://doi.org/10.1007/s00704-012-0736-5

Özdemir, O. B., Soydan, H., Çetin, Y. Y., & Düzgün, H. Ş. (2020). Neural Network Based Pavement Condition Assessment with Hyperspectral Images. *Remote Sensing 2020, Vol. 12, Page 3931*, *12*(23), 3931. https://doi.org/10.3390/RS12233931

Pan, Y., Zhang, X., Cervone, G., & Yang, L. (2018). Detection of Asphalt Pavement Potholes and Cracks Based on the Unmanned Aerial Vehicle Multispectral Imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *11*(10), 3701–3712. https://doi.org/10.1109/JSTARS.2018.2865528

Pan, Y., Zhang, X., Tian, J., Jin, X., Luo, L., & Yang, K. (2017). Mapping asphalt pavement aging and condition using multiple endmember spectral mixture analysis in Beijing, China. *Journal of Applied Remote Sensing*, *11*(1), 016003. https://doi.org/10.1117/1.JRS.11.016003

Panboonyuen, T., Jitkajornwanich, K., Lawawirojwong, S., Srestasathiern, P., & Vateekul, P. (2017). Road Segmentation of Remotely-Sensed Images Using Deep Convolutional Neural Networks with Landscape Metrics and Conditional Random Fields. *Remote Sensing*, *9*(7), 680. https://doi.org/10.3390/rs9070680

Pandey, P., Kington, J., & Kanwar, A. (2020). *Planet Basemaps for NICFI Data Program: Addendum to basemaps product specification*. *December*, 1–6.

Parfitt, J., Barthel, M., & Macnaughton, S. (2010). Food waste within food supply chains: quantification and potential for change to 2050. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1554), 3065–3081. https://doi.org/10.1098/rstb.2010.0126

Paterson, W. D. O. (1990). Road deterioration and maintenance effects. In *The highway design and maitenance standard series* (Vol. 24, p. 164). World Bank. https://doi.org/10.1016/0191-2607(90)90024-Z

Paterson, W. D. O., & Scullion, T. (1990). *Information Systems for Road Management: Draft Guidelines on System Design and Data Issues*.

Planet. (2021). *NICFI Program - Satellite Imagery and Monitoring | Planet*. https://www.planet.com/nicfi/

Potere, D. (2008). Horizontal positional accuracy of google earth's high-resolution imagery archive. *Sensors*, *8*(12), 7973–7981. https://doi.org/10.3390/s8127973

Pretorius, C. J., & Steyn, W. J. V. D. M. (2019). Quality deterioration and loss of shelf life as a result of poor road conditions. *International Journal of Postharvest Technology and Innovation*, *6*(1), 26–45. https://doi.org/10.1504/IJPTI.2019.104178

Prusky, D. (2011). Reduction of the incidence of postharvest quality losses, and future prospects. *Food Security*, *3*(4), 463–474. https://doi.org/10.1007/s12571-011-0147-y

Radoux, J., Chomé, G., Jacques, D., Waldner, F., Bellemans, N., Matton, N., Lamarche, C., D'Andrimont, R., & Defourny, P. (2016). Sentinel-2's Potential for Sub-Pixel Landscape Feature Detection. *Remote Sensing*, *8*(6), 488. https://doi.org/10.3390/rs8060488

Rajamohan, D., Gannu, B., & Rajan, K. (2015). MAARGHA: A Prototype System for Road Condition and Surface Type Estimation by Fusing Multi-Sensor Data. *ISPRS International Journal of Geo-Information*, *4*(3), 1225–1245. https://doi.org/10.3390/ijgi4031225

Rajendra Prasad, J., Kanuganti, S., Bhanegaonkar, P. N., Kumar Sarkar, A., & Arkatkar, S. (2013). Development of Relationship between Roughness (IRI) and Visible Surface Distresses: A Study on PMGSY Roads Selection and peer-review under responsibility of International Scientific Committee. *Procedia-Social and Behavioral Sciences*, *104*, 322–331. https://doi.org/10.1016/j.sbspro.2013.11.125

Rees, D., Kapinga, R., Mtunda, K., Chilosa, D., Rwiza, E., Kilima, M., Kiozya, H., & Munisi, R. (2001). Effect of damage on market value and shelf life of sweetpotato in urban markets of Tanzania. In *Tropical Science* (Vol. 41, Issue 3, pp. 142–150).

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*(5), 465–471. https://doi.org/10.1016/0005-1098(78)90005-5

Roberts, D. A., Gardner, M., Church, R., Ustin, S., Scheer, G., & Green, R. O. (1998). Mapping chaparral in the Santa Monica Mountains using multiple endmember spectral mixture models. *Remote Sensing of Environment*, *65*(3), 267–279. https://doi.org/10.1016/S0034-4257(98)00037-6

Robinson, Richard., Danielson, U., & Snaith, M. (1998). *Road maintenance management: concepts and systems*. Macmillan.

Roche. (2014). *Consultancy services for preparation of a medium-term master plan for transportation on the Volta Lake and its surrounding region*. https://new-ndpc-static1.s3.amazonaws.com/CACHES/PUBLICATIONS/2016/05/03/VOLTA+LAKE+MASTER+PLAN.pdf

Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation*. http://arxiv.org/abs/1505.04597

Rouse, J. W., Hass, R. H., Schell, J. A., & Deering, D. W. (1973). Monitoring vegetation systems in the Great Plains with ERTS. *Nasa ERTS Symposium*, 309–313.

Saana Consulting. (2016). *Diagnostic, ATWA Stage 2 Report PART 1-Corridor*. http://www.saana.com/wp-content/uploads/2015/03/ATWA-Stage-2-report_Part-I-Diagnostic_EN_4-1.pdf

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE*, *10*(3). https://doi.org/10.1371/JOURNAL.PONE.0118432

Satopää, V., Albrecht, J., Irwin, D., & Raghavan, B. (2011). Finding a “kneedle” in a haystack: Detecting knee points in system behavior. *Proceedings - International Conference on Distributed Computing Systems*, 166–171. https://doi.org/10.1109/ICDCSW.2011.20

Sayers, M. W., Gillespie, T. D., & Queiroz, C. A. v. (1986). *The International Road Roughness Experiment: Establishing Correlation and a Calibration Standard for Measurements*. World Bank.

Schnebele, E., Tanyu, B. F., Cervone, G., & Waters, N. (2015). Review of remote sensing methodologies for pavement management and assessment. *European Transport Research Review*, *7*(2). https://doi.org/10.1007/s12544-015-0156-6

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Sheahan, M., & Barrett, C. B. (2017). Food loss and waste in Sub-Saharan Africa: A critical review. In *Food Policy* (Vol. 70, pp. 1–12). Elsevier Ltd. https://doi.org/10.1016/j.foodpol.2017.03.012

Shiina, T., Umehara, H., Kaneta, T., Nakamura, N., Ito, Y., Thammawong, M., Yoshida, M., Soga, A., & Nakano, K. (2013). THE RESPONSE CHARACTERISTICS OF JAPANESE RADISH (DAIKON) TO MECHANICAL WOUND STRESS. *Acta Horticulturae*, *1005*(1005), 247–252. https://doi.org/10.17660/ActaHortic.2013.1005.27

Singh, P. P., & Garg, R. D. (2013). Automatic Road Extraction from High Resolution Satellite Image using Adaptive Global Thresholding and Morphological Operations. *Journal of the Indian Society of Remote Sensing*, *41*(3), 631–640. https://doi.org/10.1007/s12524-012-0241-4

Small, C. (2003). High spatial resolution spectral mixture analysis of urban reflectance. *Remote Sensing of Environment*, *88*(1–2), 170–186. https://doi.org/10.1016/j.rse.2003.04.008

Soleimani, B., & Ahmadi, E. (2015). Evaluation and analysis of vibration during fruit transport as a function of road conditions, suspension system and travel speeds. *Engineering in Agriculture, Environment and Food*, *8*(1), 26–32. https://doi.org/10.1016/j.eaef.2014.08.002

Suanpaga, W., & Yoshikazu, K. (2010). Riding Quality Model for Asphalt Pavement Monitoring Using Phase Array Type L-band Synthetic Aperture Radar (PALSAR). *Remote Sensing*, *2*(11), 2531–2546. https://doi.org/10.3390/rs2112531

Tadjudin, S., & Landgrebe, D. A. (2000). Robust Parameter Estimation For Mixture Model. *IEEE Transactions on Geoscience and Remote Sensing*, *38*(1), 439–445.

TCBoost, & WA Trade Hub. (2010). *West Africa Transport Logistics Analysis Using FastPath*.

The World Bank. (2020). *Trucking: A performance Assessment Framework for Policymakers*. www.worldbank.org

Thiesson, B., Meek, C., & Heckerman, D. (2001). Accelerating EM for large databases. *Machine Learning*, *45*(3), 279–299. https://doi.org/10.1023/A:1017986506241

Tröger, K., Lelea, M. A., Hensel, O., & Kaufmann, B. (2020). Re-framing post-harvest losses through a situated analysis of the pineapple value chain in Uganda. *Geoforum*, *111*, 48–61. https://doi.org/10.1016/j.geoforum.2020.02.017

Tröger, S. K. (2019). *A Systems Approach to Understanding Post-harvest Losses: Insights from a Case Study of the Pineapple Value Chain in Uganda* [DITSL]. https://doi.org/10.17170/kobra-20190313295

Tunde, A., & Adeniyi, E. (2012). Impact of Road Transport on Agricultural Development: A Nigerian Example. *Ethiopian Journal of Environmental Studies and Management*, *5*(3), 232–238. https://doi.org/10.4314/ejesm.v5i3.3

Ujuizi Laboratories. (2018). *CHEETAH Pilot project report*.

USAID. (2017). *Final Report: Study for the Development of an ECOWAS Corridor Management and Development Strategy and Action Plan*.

van Berkum, S., Dengerink, J., & Ruben, R. (2018). The food systems approach: sustainable solutions for a sufficient supply of healthy food. *Wageningen Economic Research*, *064*(June), 32. www.wur.eu/economic-research

van Wesenbeeck, C. F. A., Venus, V., Keyzer, M. A., Wesselman, B., & Asare Kye, D. (2014). *Development of a horticulture production chain in Western Africa: a case study of tomatoes in Burkina Faso and Ghana* (No. 14–01; Staff Working Paper).

Van Zeebroeck, M., Tijskens, E., Dintwa, E., Kafashan, J., Loodts, J., De Baerdemaeker, J., & Ramon, H. (2006). The discrete element method (DEM) to simulate fruit impact damage during transport and handling: Case study of vibration damage during apple bulk transport. *Postharvest Biology and Technology*, *41*(1), 92–100. https://doi.org/10.1016/j.postharvbio.2006.02.006

Wang, G. (2018). *An investigation of the suitability of smartphone devices for road condition assessment* (Issue November). https://etheses.bham.ac.uk/id/eprint/9385/7/Wang2019PhD.pdf

Wasala, W. M. C. B., Dharmasena, D. A. N., Dissanayake, T. M. R., & Thilakarathne, B. M. K. S. (2015). Vibration simulation testing of banana bulk transport packaging systems. *Tropical Agricultural Research*, *26*(2), 355. https://doi.org/10.4038/tar.v26i2.8098

Williams, P. M. (1995). Bayesian Regularization and Pruning Using a Laplace Prior. *Neural Computation*, *7*(1), 117–143. https://doi.org/10.1162/neco.1995.7.1.117

Winter, M. E. (1999). N-FINDR: an algorithm for fast autonomous spectral end-member determination in hyperspectral data. In M. R. Descour & S. S. Shen (Eds.), *Imaging Spectrometry V* (Vol. 3753, pp. 266–275). SPIE. https://doi.org/10.1117/12.366289

Workman, R., Otto, A., & Irving, A. (2016). *The use of appropriate high-tech solutions for road network and condition analysis, with a focus on satellite imagery* (Issue GEN2070A).

World Bank. (2011). Missing food: the case of postharvest grain losses in sub-Saharan Africa. In *World Bank* (Issues 60371-AFR). World Bank.

World Bank. (2019). *Enhancing Burkina Faso Regional Connectivity: An Economic Corridor Approach*. World Bank, Washington, DC. https://doi.org/10.1596/33040

World Economic Forum. (2017). *The Global Competitiveness Report 2017-2018*.

Xu, Y., Xie, Z., Feng, Y., & Chen, Z. (2018). Road Extraction from High-Resolution Remote Sensing Imagery Using Deep Learning. *Remote Sensing*, *10*(9), 1461. https://doi.org/10.3390/rs10091461

Yang, X., Qin, Q., Grussenmeyer, P., & Koehl, M. (2018). Urban surface water body detection with suppressed built-up noise based on water indices from Sentinel-2 MSI imagery. *Remote Sensing of Environment*, *219*, 259–270. https://doi.org/10.1016/j.rse.2018.09.016

Zare, A. (2008). *Hyperspectral endmember detection and band selection using bayesian methods*.

Zare, A., & Gader, P. (2007). Sparsity Promoting Iterated Constrained Endmember Detection in Hyperspectral Imagery. *IEEE Geoscience and Remote Sensing Letters*, *4*(3), 446–450. https://doi.org/10.1109/LGRS.2007.895727

Zha, Y., Gao, J., & Ni, S. (2003). Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *International Journal of Remote Sensing*, *24*(3), 583–594. https://doi.org/10.1080/01431160304987

Zhang, Z., Liu, Q., & Wang, Y. (2017). Road Extraction by Deep Residual U-Net. *IEEE Geoscience and Remote Sensing Letters*. https://doi.org/10.1109/LGRS.2018.2802944

Zhao, B., Zhong, Y., Ma, A., & Zhang, L. (2016). A Spatial Gaussian Mixture Model for Optical Remote Sensing Image Clustering. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*(12), 5748–5759. https://doi.org/10.1109/JSTARS.2016.2546918

Zhao, Q., Hautamaki, V., & Fränti, P. (2008). Knee Point Detection in BIC for Detecting the Number of Clusters. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 5259 LNCS* (pp. 664–673). https://doi.org/10.1007/978-3-540-88458-3_60