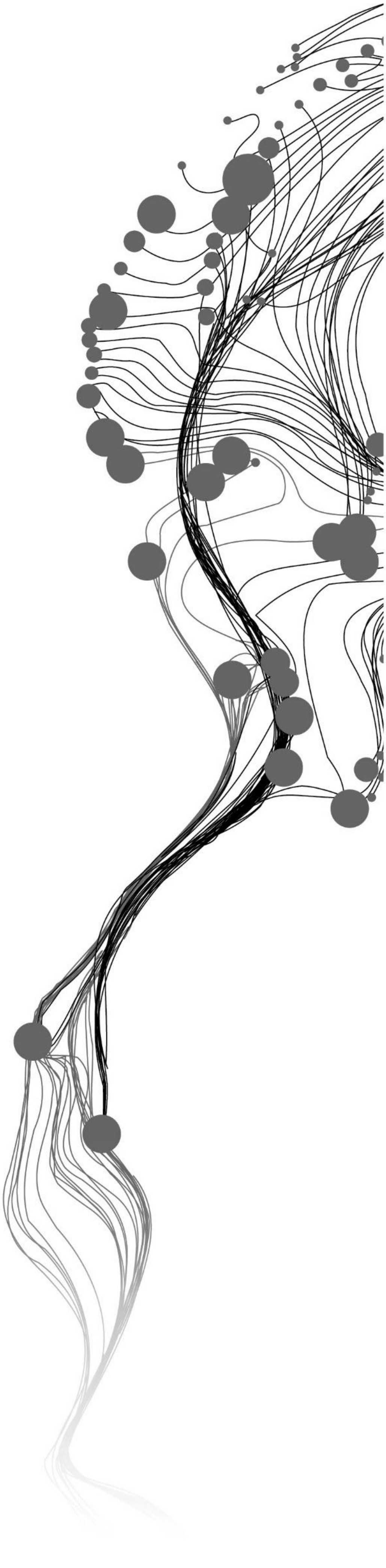
INTEGRATING REMOTE SENSING AND STREET VIEW IMAGES TO MAP SLUMS USING DEEP LEARNING APPROACH

ABBAS NAJMI
August, 2021

SUPERVISORS:
Prof. dr. Richard. V. Sliuzas
Dr. Caroline. M. Gevaert

ADVISORS:
Dr. Divyani Kohli
Dr. Monika Kuffer



INTEGRATING REMOTE SENSING AND STREET VIEW IMAGES TO MAP SLUMS USING DEEP LEARNING APPROACH

ABBAS NAJMI
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 Geo-information Science and Earth Observation.
Specialization: Urban Planning and Management

SUPERVISORS:
Prof. dr. Richard. V. Sliuzas
Dr. Caroline. M. Gevaert

ADVISORS:
Dr. Divyani Kohli
Dr. Monika Kuffer

THESIS ASSESSMENT BOARD:
Dr. Javier. A. Martinez (Chair)
Dr. Taïs Grippa (External Examiner, Université Libre de Bruxelles)

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

The United Nations includes slum upliftment as one of the agenda in the Sustainable Development Goals 11, Target 11.1- "safe and affordable housing" to fight against poverty. The information to keep track of target 11.1, such as physical location and size of slums, is lacking or inadequate in governmental documents. Therefore it is vital to map slums in order to comprehend the existing situation and build future slum development policy plans to achieve target 11.1. Remote Sensing (RS)-based approaches have gained much recognition in the slum mapping field in the last few decades due to the availability of Remote Sensing Imagery (RSI) of Very High Resolution (VHR). In RS-based approaches, the Deep Learning (DL) approaches such as Fully Convolutional Network (FCN) have been shown to achieve reasonably higher accuracies for slum mapping than other RS-based approaches. However, using RSI alone has its limitation, i.e., the absence of ground-level information, making slum identification difficult in the dense urban scene. Previous studies show that adding ground-level information with RSI can help identify slums more precisely than using RSI alone, but none of the studies used Street View Imagery (SVI) as the source of ground-level information to compliment RSI in the field of slum mapping. Therefore this research aims to integrate RSI with SVI using FCN for slum mapping. Implementing FCN has three significant challenges, from which the first challenge is general for all slum mapping approaches, and the remaining two are specifically for the FCN. First is the conceptualization of slums because there is no unique definition of slums, i.e., it varies from institution to institution. Second, extraction of ground-level information through SVI to identify slums. Third, setting up an FCN pipeline to integrate overhead information with ground-level information, i.e., integrating RSI with extracted features of SVI.

The city of Jakarta was chosen for this study because of two main reasons. First, the presence of kampungs (urban villages) in Jakarta. Around 60% of Jakarta's population lives in kampungs, and the diverse socioeconomic conditions in kampungs make it challenging to identify slums inside kampungs, i.e., the line between slums and non-slums is vague. There are two types of kampungs such as legal and illegal. This research focused on the illegal kampungs called slums. Second, there are various local definitions of slums used in Jakarta, making the conceptualization of slums more difficult. Initially, the western region of Jakarta was chosen for study because of the high density of slum settlements according to the official slum reference map of 2017, but due to data constraints, approximately half of the western region with some part of the northern and central region was selected as a study area.

In this research, four deep neural networks are applied with different datasets, i.e., FCN-DK6 used RSI alone, Places365-VGG16 was fine-tuned using SVI, and FCN-DK6-i and Modified FCN-DK6 used a combination of RSI and SVI in the study area. The FCN-DK6 network was trained with RSI alone to map slums in the study area. The Places365-VGG16 network was fine-tuned in the context of Jakarta's slums using SVI captured in the study area. Further, the fine-tuned Places365-VGG16 network was used to extract the features from widely dispersed SVI and spatially interpolated them to precisely match the spatial resolution of RSI, which are combined with RSI for slum mapping using FCN-DK6-i and Modified FCN-DK6 networks. The result shows that the Modified FCN-DK6 outperforms FCN-DK6 and FCN-DK6-i in slum mapping, demonstrating that combining RSI and SVI can achieve higher accuracy because SVI contains useful ground-level information which helps to identify slums in an urban setting than using RSI alone. Furthermore, we describe experimental investigations by combining the extracted SVI features with RSI at different levels in FCN-DK6-i and Modified FCN-DK6, which shows that the combination of RSI and SVI can improve the accuracy obtained from RSI alone, but it also depends on how they are integrated. The Modified FCN-DK6 presented here obtains better results than a direct integration through FCN-DK6-i.

Keywords: Slum mapping, Remote sensing images, Street view images, Deep learning, Fully convolutional networks, Feature extraction, Spatial interpolation

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to my supervisors, Prof. dr. R.V. Sliuzas and Dr. C.M. Gevaert for accepting my research topic at the last moment of the research proposal phase and for providing invaluable support with constructive feedback throughout the research process. I would also like to express my profound gratitude to my advisors Dr. D. Kohli & Dr. M. Kuffer, for supporting me throughout my research phase by providing me the valuable input and for helping me in selecting the study area.

My deepest gratitude to Mr. Jati Pratomo, MSc., for being my local knowledge expert and for providing valuable inputs for understanding the context of slums in Jakarta to generate the slum reference map, which was the crucial step for my research. My special thanks to European Space Agency (ESA) for accepting my project proposal and providing me with the GeoEye Image dataset to support my research. I am grateful for their valuable assistance. I would like to thank all the academic staff at the PGM Department, Faculty of ITC, the University of Twente for their sharing knowledge, which will be an added asset in my future profession. I would like to thank the CRIB coordinator Dr. Serkan Girgin for helping me with the CRIB platform issues and supporting my work with Deep learning architectures. I would like to acknowledge the Google Map platform, which provided me with the necessary google street view images for my study. I would also extend my thanks to the open-source community for contributing to the development of open-source software, programming languages, and libraries. I want to thank everyone involved in the development of TensorFlow, Keras, QGIS, and MSOffice.

I want to use this moment to convey gratitude to my parents Mr. Khuzaima Najmi, Mrs. Khurshida Najmi, and my brother's family, Dr. Hussain Najmi, Mrs. Zainab Najmi, Ms. Sara Najmi, and I can't thank them enough for their love and support during my journey so far. I would not be here without their support. I want to extend my gratitude to my mentor from ISRO, Dr. Ashok Kumar Joshi and Dr. Abdul Hakeem Kaja Mohideen, who showed me the way to ITC for my master's and our family friend Dr. Hemant Mahiyar for encouraging me to do master.

I am grateful to my friends Praneeth, Sowmeya, Mahesh, Dhananjay, Darshana, Vijayudu, Ranit, Harsh, Sidhant, and fellow students in UPM who made this master's journey memorable, and I hope to collaborate with them in professional careers in the future.

Enschede, August 2021

Abbas Najmi

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

SDG	: Sustainable Development Goal
MDG	: Millennium Development Goal
RS	: Remote Sensing
RSI	: Remote Sensing Imagery
VHR	: Very High Resolution
ML	: Machine Learning
OBIA	: Object-Based Image Analysis
RF	: Random Forest
SVM	: Support Vector Machine
DL	: Deep Learning
FCN	: Fully Convolutional Network
SVI	: Street View Images
CNN	: Convolutional Neural Network
FC	: Fully Connected
PPV	: Positive Prediction Value
IoU	: Intersection over Union
CBD	: Central Business District
ESA	: European Space Agency
PCA	: Principal Component Analysis
IDW	: Inverse Distance Weighted
OA	: Overall Accuracy

1. INTRODUCTION

1.1. Background and Justification

Urbanization is a global megatrend that is unstoppable and irreversible (United Nations-Habitat [UN-Habitat], 2018). More than half of the population currently live in urban areas in this rapidly urbanizing world and is expected to increase to 68% by 2050 (United Nations Department of Economic and Social Affairs [UNDESA], 2018). Rapid urbanization and inadequate city planning increase pressure on necessary infrastructure and services such as lack of affordable housing, sanitation, water, waste management, and roads, which leads to increased slums and slum dwellers. According to the United Nations Department of Economic and Social Affairs [UNDESA] (2020), more than one billion people currently live in slums or informal settlements. Most of these informal settlements' growth has happened in developing regions such as Northern Africa, Western Asia, sub-Saharan Africa, and South Asia (UNDESA, 2020). These regions have limited resources and capacity to overcome development challenges that result in unplanned urbanization. These unplanned urbanization areas promote informal settlements' growth, resulting in urban poverty, inadequate housing, and inequality (UN-Habitat, 2018).

In the last two decades, the reduction of informal settlements or slums has been a high priority on the worldwide agenda. In the year 2000, a goal has been set to uplift at least 100 million slum dwellers by the end of 2020 under Millennium Development Goal (MDG)-7 (United Nations Development Programme [UNDP], 2016). In contrast to the MDG-7 target, 320 million slum dwellers were uplifted, i.e., gained access to basic amenities such as drinking water, sanitation, and less populated dwellings between 2000 and 2014, exceeding the set target (UNDESA, 2020). Further in 2015, the new framework has been proposed with 17 different goals under Sustainable Development Goals (SDG) for 2030 (United Nations Department of Economic and Social Affairs [UNDESA], 2015). The global slum reduction goal is addressed under **SDG 11**- "Make cities and human settlements inclusive, safe, resilient and sustainable," **Target 11.1**-"Safe and affordable housing" (United Nations Department of Economic and Social Affairs [UNDESA], 2017, p. 11). The goal of Target 11.1 is to "ensure access for all to adequate, safe, and affordable housing and basic services and upgrade slums" (UNDESA, 2017, p. 11).

According to UN-Habitat (2018), slums and informal settlements have significant overlap in terms of physical characteristics, but some informal settlements may have good living conditions and even be fairly wealthy. On the other hand, the settlement is called a slum if at least one of the following criteria is fulfilled: (1) absence of tenure security, (2) lack of housing durability, (3) insufficient living spaces, (4) lack of access to water and sanitation (UN-Habitat, 2018) and these criteria are used to identify slums in urban environment. Slums are quite dynamic, i.e., slum characteristics change over time, such dense structure, location, building size and height, and building arrangement, making slum identification extremely complex. Different indicators are used to understand the complexity of slum areas in the urban scene on a local level (Kohli, Kerle, and Sliuzas, 2012). The government continuously improves the existing situation by constructing and implementing pro-poor policies (Arimah, 2011), providing necessary infrastructure and amenities to uplift slum dwellers from their current conditions. Generally, the spatial information regarding slum areas is missing or incomplete from the official records (Nijman, 2008); hence, it is necessary to identify slum areas to understand the current situation for further slum development policy plans (Duque, Patino, Ruiz, and Pardo-Pascual, 2015).

Slum mapping is complicated because it is extremely difficult to define the actual boundary of slums in an urban environment. The process of slum mapping involves different stakeholders like government, private, and the public in various disciplines such as economic and social environments. The stakeholders must understand the different levels of slum and their characteristics to produce a slum map. A slum map is an efficient way to express the spatial distribution and information about slums helps governmental organizations to make better decisions for slum upgrading plans and policies.

There are three approaches for mapping slums: survey-based approaches, participatory approaches, and Remote Sensing (RS) based approaches (Mahabir, Croitoru, Crooks, Agouris, & Stefanidis, 2018). Survey-based approaches contain long temporal gaps between the data collections, and it is also time and resource-intensive. However, they are still very useful in some cases where ground data is needed, such as population statistics (Kohli et al., 2012). Often slums have been ignored in these formal surveys while collecting mapping data (Joshi, Sen, and Hobson, 2002). In participatory approaches, local people are involved in making a better perception of reality, but there can be a conflict of interest between people's perceptions which may lead to different results. For example, one person uses a lack of access to water and sanitation as an indicator to identify slums, but maybe the other person won't use the same indicator; thus, both persons have different perspectives to identify slums. Participatory approaches also take lots of resources, time, and money to implement, and the data obtained from this approach can be highly accurate because they collect data on the ground (Kohli et al., 2012). RS-based approaches reduce human effort and time but need an RS expert to analyze the data. RS data helps to analyze the situation in real-time (Hofmann, Strobl, Blaschke, and Kux., 2008) and provides up-to-date information with a birds-eye view, including areas with no available data.

In the last few decades, the RS approach gained a lot of recognition in the research community with the large availability of Very High Resolution (VHR) Remote Sensing Imagery (RSI) (Kuffer, Pfeffer, & Sliuzas, 2016). Researchers have developed different RS-based approaches for slum mapping; the primary step for most approaches is to define and design different sets of criteria through which slum and non-slum can be differentiated from RS imagery (Mahabir et al., 2018). However, different RS-based approaches of slum mapping are challenged with varying morphological features and characteristics of slums within and across the cities (Kuffer et al., 2016). This complexity makes the designed criteria limited to those specific areas only with the unique dataset usage. If the designed criteria are used with some other dataset (imagery) or different areas within the city boundary, they might perform poorly because of different morphological features. In such cases, Machine Learning (ML) approaches outperform the classical RS-based approaches, such as Object-Based Image Analysis (OBIA) for slum mapping (Kuffer et al., 2016). ML approaches extract spatial features by long-range pixels from RSI to map slums (Persello & Stein, 2017). Thus, ML-based approaches such as Random Forest (RF) and Support Vector Machine (SVM) will produce better results than the classical RS-based approach. Still, ML approaches required a clear notion of slum characteristics (Leonita, Kuffer, Sliuzas, and Persello, 2018). As stated above, a proper understanding of local and contextual knowledge of slum is required because there is no universal conceptualization of slums, i.e., the definition of slum changes with area and time, and it is highly dependent on the local or national governmental bodies.

In contrast to traditional ML approaches such as RF and SVM, Deep Learning (DL) approaches such as Fully Convolutional Network (FCN) consist of different stack layers that help extract more accurate information from input imagery to identify slum areas with higher accuracy (Persello & Stein, 2017; Hoeser & Kuenzer, 2020). Different studies show that FCN can be used for slum mapping through RSI (Ajami, Kuffer, Persello, and Pfeffer, 2019). However, researchers could not fully understand the complexity of urban forms to map slums by using RSI. The limitation of using only RSI is the absence of ground-level information such as inferior building materials, open drainages, and the number of floors. The ground-level information can be inferred through ground surveys, interviews, and Street View Images (SVI), i.e., street-level photographs. Only a few studies have been carried out to communicate the integration of different ground-truth dataset with RSI to delineate slums in urban scene. In comparison, none of the researchers used SVI to compliment RSI for mapping slums.

Previous studies used RSI and SVI individually to map or identify slums in a dense urban scene. For example, Ibrahim, Haworth, and Cheng (2019) use only SVI to identify slums using SlumNet architecture based on the Convolutional Neural Network (CNN) model, recognized the difference between slum or non-slum urban scenes. The architecture of SlumNet consists of 10 hidden layers in which two are fully connected layers. SlumNet did not accurately classify slums or non-slum due to little understanding of the urban scene. The author did not correctly conceptualize slums and downloaded random slum images of Africa and Egypt from the internet to fine-tune the pre-trained model, as we know the morphological characteristics of slums vary with places due to which the model did not perform well. There is always a possibility of error that exists while mapping slums using RSI and SVI individually. The combination of RSI and SVI can be potentially used to quantify slums more precisely as it combines the bird-eye view of the VHR images and the ground images (SVI) with additional feature information. Thus, this study explores the potential of integrating SVI with VHR satellite imagery using state-of-the-art deep learning algorithms to map slums in the dense urban scene.

1.2. Research Gap and Innovation Point

Several studies have shown that the physical characteristics of slums can be examined using VHR satellite imagery for slum mapping via visual image interpretation, OBIA, and ML approaches. The ML approach shows remarkable performance in slum mapping because it incorporates spatial, spectral, textural, and structural features (Kuffer et al., 2016). However, slum mapping is difficult using RSI alone because the RSI captures the urban environment from a bird-eye view, resulting in a lack of ground-level information, which plays a crucial role in slum mapping. Nowadays, the increased open-source of geotagged data can help us infer the ground-level information that can be further combined with RSI for slum mapping. For example, SVI can be used for accessing ground-level information.

Previously researchers have used RSI alone to map slums, whereas only very few researchers have used SVI to identify slums. There is always a possibility of misclassification in mapping slums using RSI because the information in RSI is limited to overhead information. For example, it might be possible that slum and non-slum areas share the same physical characteristics like high building density, which can cause misleading results because the human eye may not significantly recognize the features captured through RSI. In contrast, using SVI alone can help identify slum and non-slum settlements, but the slum map can not be generated because SVI are captured along the road with limited coverage around the point from which it is captured. For example, if we want to map an area that is only accessible by foot, those areas can not be map through SVI because SVI are taken in those areas accessible by motorbike or car.

The complementary information from SVI can be used with the RSI to understand the complex urban scene, and it can be hypothesized that the combination of RSI with SVI may lead to better results in slum mapping. As mentioned in the above literature, none of the researchers integrated the ground-level information extracted from SVI with RSI to map slum areas in the dense urban scene using the DL approaches.

1.3. Research Objective and Questions

1.3.1. Main objective

This study aims to integrate remote sensing images and street view images using a deep learning model to map slums in the complex urban scene of Jakarta, Indonesia.

1.3.2. Sub objectives and Research Questions

I. To identify the characteristics of slums versus non-slum in the study area.
- What are the physical characteristics of slums in the study area?
- Which features can be extracted from RSI to classify slums?
- Which visual features can be extracted from SVI to classify slums?

II. To incorporate SVI with RSI for slum mapping using FCN.
- Which FCN architecture is the best fit for using the combination of RSI and SVI to identify slums?
- What is a suitable grid size?
- Which technique can be used to interpolate the feature vector of SVI into the 2-dimensional space of RSI?
- How to deal with the incomplete data of SVI?

III. To investigate the significance of using SVI for mapping slums.
- What is the added value of combining SVI and RSI for mapping slums?

1.4. Research Conceptual Framework

As previously stated in Section 1.1, rapid urbanization makes it very challenging for the government to develop and enforce effective city planning and puts extensive pressure on essential infrastructure and services. Therefore we need to know which areas in the city are deprived in terms of essential services so that the government can make policy to uplift those areas. Different approaches use ground-level data (SVI) or overhead data (RSI) to delineate slum areas, as discussed in Section 1.2. However, RSI and SVI contain complementary information. Therefore we propose an innovative method to integrate two different datasets for mapping slums. Figure 1.1 shows the conceptual framework of this research.

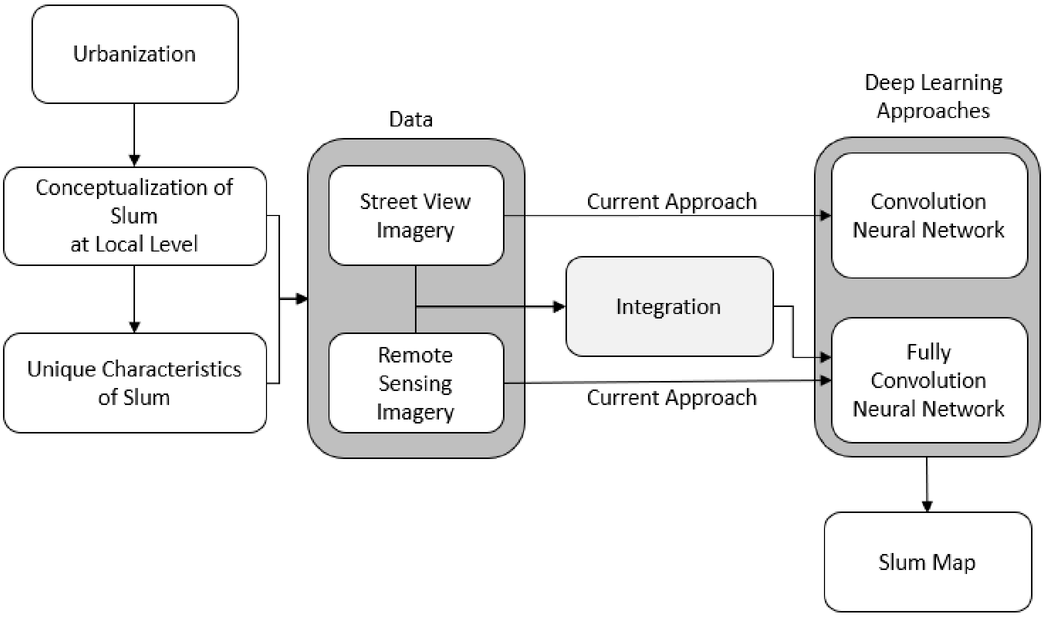


Figure 1.1: Research Conceptual Framework

1.5. Thesis Structure

The thesis is divided into different chapters. Chapter 2 provides a detailed literature review to understand the different slum mapping approaches that evolved in the last few decades, mainly focusing on the deep learning approach. Chapter 3 describes the study area and discusses slum dynamics and characteristics of slums in Jakarta. Chapter 4 provides a detailed description of the datasets used in this research. Chapter 5 explains the detailed methodology to achieve the research objective by answering the research questions. Chapter 6 presents the outcome of the research. Chapter 7 provides a detailed discussion on the research outcome. Finally, Chapter 8 summarizes the research by presenting the research's conclusions and limitations and suggests recommendations for future work.

2. LITERATURE REVIEW

This chapter reviews the literature and illuminates the direction of this research work. It starts with Section 2.1 by reviewing various literature on slums for understanding how slums are conceptualized in different research articles. Section 2.2 reviewed different approaches for slum mapping in the domain of RS. Finally, section 2.3 reviewed different DL approaches and accuracy matrices for slum mapping, which was further used in this research.

2.1. Complexity in Defining Slums

Different terminology has been used in literature to refer slums, such as "informal," illegal," "squatter," "irregular," "unplanned," "deprived," or "substandard settlement/area" (Kuffer et al., 2016, p. 6). These terms have been used interchangeably with slums by different authors (Kuffer et al., 2016).

Slums do not have any universal definition (Verma, Jana, and Ramamritham, 2019). However, United Nations has defined slums on a broader scale, as mentioned in Section 1.1, but due to the varying characteristics of slums such as lack of basic service and infrastructure (e.g., electricity, sanitation, water), overcrowding, construction materials, hygiene and health, crime and violence, land tenure and security, etc. (United Nations-Habitat [UN-Habitat], 2003), it is hard to address slums with one unique definition; therefore, the definition of slums can vary in different regions. Generally, the definition of a slum depends on different standards of local or national government authorities, and these authorities conceptualize slums differently. For example, the Bangkok government uses overcrowding, health and hygiene, crime and violence, and surrounding environment indicators to define slum areas (UN-Habitat, 2003). Most South Asian countries use insecure land tenure, lack of access to water and sanitation, and overcrowding as major indicators to define slum areas (UN-Habitat, 2003).

According to Lilford et al. (2019), slums can be conceptualized in two ways. The first approach is "feature first," which generally depends on household-level surveys. According to local or national authorities' standards, the observed features of slums and non-slums are identified first. Then the area is defined based on the observed features; therefore, it is also called a bottom-up approach. The second approach is the "space first" or top-down approach because it starts with selecting an area first. Then the selected area is classified into slum and non-slum based on features.

According to Kuffer et al. (2016), there are various physical characteristics of slums, which differentiate slums from non-slum built-up areas, such as small roof size, high roof coverage density, poor building materials, smaller and irregular building size. However, some physical features, such as building density and building size, can be delineated using RSI, but physical features like poor building materials can not be identified using RSI. The measurement of physical features can be problematic using RSI alone even if they are appropriately defined (Pratomo, Kuffer, Kohli, and Martinez, 2018); these problems can arise due to a lack of local contextual knowledge to conceptualize slums (Kuffer et al., 2016). For example, some historical settlements can be easily misclassified as slums because they have the same morphological characteristics as slums (Kuffer et al., 2016). Thus the resembling physical characteristics of slum and non-slum area makes it more uncertain about using RSI alone.

2.2. Remote Sensing-Based Slum Mapping

In the last few decades, several approaches have been developed to map slums with VHR images (Mahabir et al., 2018). The slum mapping approach can broadly be divided into three types: visual image interpretation, OBIA-based approaches, ML-based approaches (Mahabir et al., 2018).

Visual image interpretation can map slums with quite a reasonable accuracy rate (Taubenböck & Kraff, 2014). The visual image interpretation approach is time-consuming and has some uncertainties with boundary delineation because it depends on how the interpreter perceives slums (Pratomo et al., 2018). Generally, the data mapped using visual image interpretation is used as reference data for cross-checking the results from other approaches.

OBIA is a popular approach to map slums (Kuffer et al., 2016). The image is divided into different meaningful objects with their geographic information, and then the characteristics of those objects are computed (Blaschke et al., 2014). OBIA outperforms the conventional pixel-based approaches because OBIA handles the input images as a set of objects instead of pixels and integrates different spatial, spectral, and contextual properties of the selected object for classification (Kohli, Warwadekar, Kerle, Sliuzas, and Stein, 2013). In contrast, the pixel with the same reflectance is assigned to the same class in pixel-based classification. The common problem with the pixel-based classification is the salt and pepper effect because it relies only on an object's spectral signatures (Kohli et al., 2013). Generally, OBIA is used with VHR satellite imagery. However, the concept of slum should be clear while defining the set of rules for OBIA for slum mapping (Kohli et al., 2013). Kohli et al. (2013) map slums using OBIA in Ahmedabad, tested the accuracy using different datasets and achieved overall accuracy ranging from 47 to 68%. The accuracy of OBIA decreases with the increase of urban environment complexity, i.e., sometimes the roofing material of slum and non-slum show the same spectral reflectance, making it hard to capture the characteristics of slums (Kuffer et al., 2016). Thus to overcome misclassification from OBIA, the OBIA ruleset can be combined with ML approaches such as Support Vector Machine (SVM) (Zahidi, Yusuf, Hamedianfar, Shafri, and Mohamed, 2015).

In general, ML-based approaches perform better than classical RS-based classification approaches (Verma et al., 2019). ML-based approaches are frequently used for slum mapping, and these approaches are data-driven, i.e., heavily dependent on a large amount of data. The availability of large data set with extensive pixel-based information makes ML approaches ideal for image classification (Verma et al., 2019). Duque, Patino, and Betancourt (2017) and Kuffer et al. (2018) explore ML algorithms such as Random Forest (RF) and SVM for slum mapping; the SVM achieved an F1 score varying between 0.73 to 0.92, and RF achieved F1 score varies between 0.72 to 0.94. Previous studies show that the ML approaches produce a better result than classical RS-based approaches. However, ML approaches' accuracy depends on feature selection, requiring a clear understanding of the local contextual knowledge of slums (Leonita et al., 2018). ML algorithm learns features from the training data set to generate output from the unknown input (Persello & Stein, 2017; Hoeser & Kuenzer, 2020). In contrast, DL consists of different stack layers that help extract more accurate information from input data.

DL is part of ML, popular in the scientific community for slum mapping because high accuracy can be achieved using the DL approach (Kuffer et al., 2016). DL recently gained attention in the RS community due to the open-source DL models (Verma et al., 2019). DL consists of CNN and FCN. CNN is one of the main image classification approaches in DL, and FCN is derived from classical CNN. Currently, CNN and FCN are getting attention for mapping slums (Persello & Stein, 2017).

DL model extracts the information from the input data using different convolutional layers and further predicts and displays the result using the classification layer. DL consists of more than two layers (Zhu et al., 2017). According to the training data set, the weights are optimized to different layers and reduce the prediction error (Persello & Stein, 2017). Therefore it is irrelevant to design the rule set or selection of features for the classification. DL approaches can be used in the complex urban environment with different dataset combinations, i.e., a combination of overhead data and ground-level data. However, the results of the DL model heavily rely on reference data (ground truth data) used for training.

2.3. Deep Learning-Based Approach

2.3.1. Convolutional Neural Networks

CNN is a patch-based classification approach. The CNN classifies the input images' central pixels and labels them accordingly in the output (Michael, Neal, Burke, Lobell, & Ermon, 2016). CNN architecture includes feature extraction and classification layers. Generally, CNN consists of four layers: convolutional layers, non-linear activation, pooling layers, and fully connected (FC) layers. All the layers are trained throughout the network. Figure 2.1 shows the CNN architecture acquired from Mboga, Persello, Bergado, and Stein (2017). A standard CNN consists of multiple convolutional and fully connected layers. The convolutional layer extracts the image features from the training dataset and converts them into a one-dimensional array vector, further given as input to the FC layer. Then the output from the FC layer is passed through to the activation layer (softmax) for image classification. Thus, both convolutional and FC layers are accountable for learning classification rules (Persello & Stein, 2017).

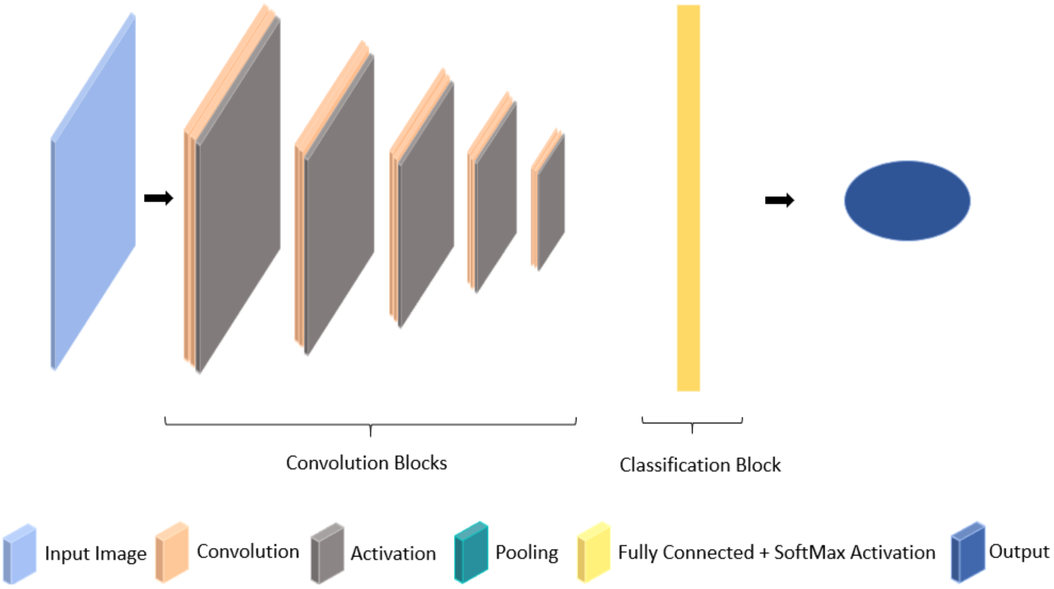


Figure 2.1: CNN architecture acquired from Mboga et al. (2017)

Researchers have found that CNN can outperform previous RS-based approaches (Persello & Stein, 2017). Verma et al. (2019) used CNN to map slums in Mumbai. The author obtained overall accuracy and kappa coefficient of about 94.2 % and 0.70 for VHR imagery and 90.2 % and 0.55 for Medium resolution (MR) imagery. Mboga et al. (2017) used CNN to map slums in Dar es Salaam with Quick bird imagery and obtained an overall accuracy of 91.71%. Michael et al. (2016) used CNN with nighttime satellite imagery to map slums in African countries.

In contrast to RSI, Ibrahim et al. (2019) used VGG16 CNN to identify slums using SVI and achieved the validation accuracy of 85%, but the model did not perform well in the complex urban environments, i.e., in those areas where slum and non-slum have similar characteristics. There are two key barriers in implementing CNN to a large RSI or aerial dataset: (i) a large amount of reference data is needed, and (ii) high computation costs (Persello & Stein, 2017) because of which the learnable parameters become larger than the convolutional layers while training. FCN is derived from CNN, which might overcome the barrier of CNN in terms of high computational cost.

2.3.2. Fully Convolutional Neural Networks

FCN is a pixel-based classification approach and is also trained throughout all the layers. It is also called a semantic segmentation network. In FCN, deconvolutional layers replace the FC layers, allowing flexibility in the input size. The convolutional–deconvolutional layer or dilated kernel layer helps keep the output similar to the input in terms of size and resolution of an input image (Long, Shelhamer, and Darrell, 2015; Wurm, Stark, Zhu, Weigand, and Taubenböck, 2019), resulting in the lower computational cost of FCN compared to CNN. The result in the FCN requires less computational cost than CNN.

FCN consists of five parts: (1) convolutional layers; (2) non-linear activation functions (e.g. leaky Rectified Linear Unit (lReLU)); (3) pooling (e.g. max pooling); (4) deconvolutional layers; (5) classification layers (e.g. Softmax). The deconvolutional layer improves the model performance and reduces the chances of overfitting (Teerapong, Kulsawasd, Siam, Panu, and Peerapon, 2017). Figure 2.2 shows the Encoder-Decoder FCN architecture acquired from Teerapong et al. (2017).

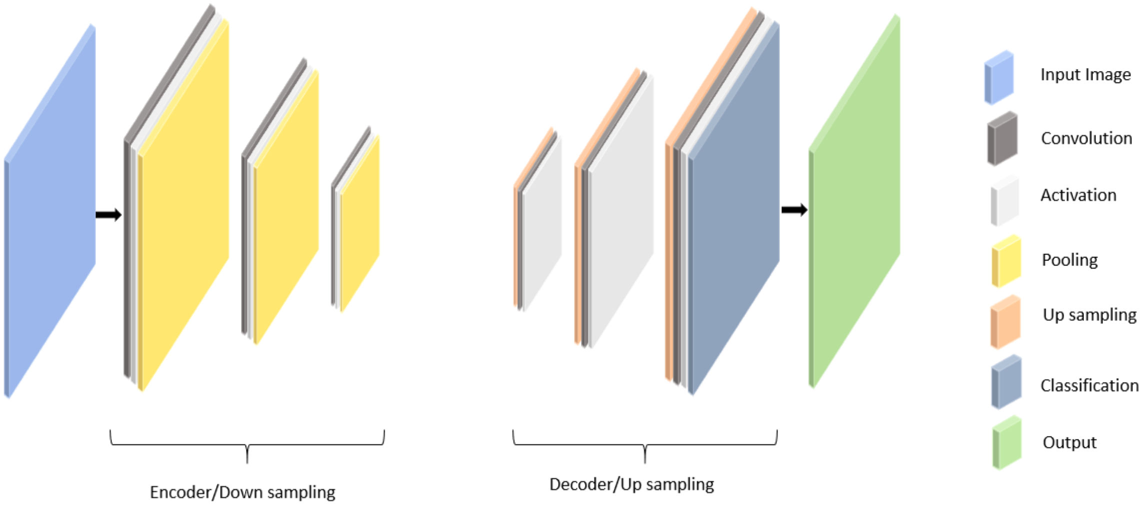


Figure 2.2: Encoder-Decoder FCN architecture acquired from Teerapong et al. (2017)

Few studies use FCN for slum mapping. Wurm et al. (2019) explored the FCN-VGG19 to identify slums in Mumbai using different sensors for slum mapping. The model was trained on QuickBird imagery and obtained 86% of Positive Prediction Value (PPV). Further, the trained model is transferred to different datasets such as Sentinel -2 and TerraSAR-X and achieved 38% and 79% PPV. Stark et al. (2019) explored FCN-VGG19 with pre-trained weights from Imagenet and fine-tuned it to identify slum areas in Mumbai and Delhi. The author achieved an accuracy of 64% for Mumbai and 34% for Delhi because slum structures of Mumbai differ from Delhi, and slums and non-slum areas of Delhi make the transfer learning a bit difficult (Stark et al., 2019).

FCN also uses dilated kernels technique to increase the sizes of the receptive fields (RFs) (Persello & Stein, 2017). The FCN-architectures with dilated kernel technique do not have deconvolutional layers, and it is called FCN-DK (fully connected convolutional neural network with the dilated kernel) (Persello & Stein, 2017).

FCN-DK reduces the number of features that prevent the overfitting of data and lower the computational cost compared to other FCN networks. FCN-DK consists of different convolutional blocks. Each convolutional block comprises four layers: zero-padding layers, convolutional layers (different dilated rates in separate blocks), activation layers, and pooling layers, which are finally connected to the classification layer. Figure 2.3 shows the FCN-Dk3 architecture proposed by Persello & Stein (2017).

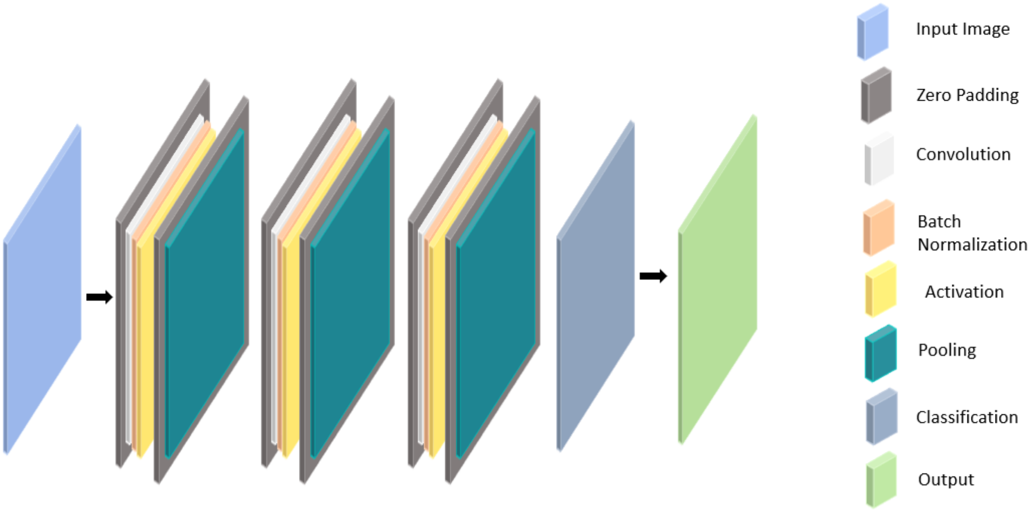


Figure 2.3: FCN-DK3 architecture proposed by Persello and Stein (2017)

Persello & Stein (2017) compared the performance of CNN, SVM, and different FCN-DKs such as FCN-DK3, FCN-DK4, FCN-DK5, and FCN-DK6 for slum mapping in which the FCN-DK6 outperformed other models with an overall accuracy (OA) of 84%. However, the accuracy of FCN-DKs will be reduced if applied in the complex urban environment where there is an overlap between the feature of slum and non-slum. These networks extract the features based on the input dataset, i.e., if slum and non-slum don't have the distinct feature on the satellite imagery, the result will probably not be satisfactory. Therefore, we need to simplify the urban complexity by using an additional dataset with RSI, which will help the network to understand the urban environment better so that slum and non-slum can have distinct features.

2.3.3. Techniques for Training Deep Learning Network

DL model can be trained in two ways. The first option is to adjust, i.e., fine-tune the pre-trained network to match the current classification requirement. Thus, the pre-trained network effectively reduces the required training data and computational cost because the network was already trained on the generalized dataset to classify the required class. In some cases, the generalized dataset consists of a somewhat similar class, not the same class on which it is further fine-tuned. At the time of fine-tuning, some of the initial convolutional layers can be frozen, and a new trainable convolutional layer can be added after the frozen layers because the pre-trained model has a broader understanding of features of the required class. The second option is to train the DL model from scratch, but it requires intensive training data and higher computational requirements, resulting in lesser accuracy. For example, Stark et al. (2019) set up two experimental models (i) fine-tune the pre-trained FCN-VGG19 on ImageNet (ii) train FCN-VGG19 from scratch. As a result, the first model produces an accuracy of 69%, whereas the second model produces an accuracy of 34% only.

2.3.4. Multimodal Data Fusion

Recently, researchers are integrating macro overhead data (e.g., Satellite images) and micro ground-level data to understand urban environment better (Cao et al., 2018). Researchers have explored the combination of different data sources using the DL approach, such as Zhao et al. (2019) identified the geographical object using high-resolution (HR) RSI and OpenStreetMap (OSM). Recently few authors integrated the RSI and SVI to map different urban land-use/land cover using CNN and FCN. Workman, Zhai, Crandall, and Jacobs (2017) used kernel regression and density estimation technique to convert the extracted features from the SVI to generate the dense feature maps and further interpolate them with the

RSI using Nadaraya–Watson kernel regression technique. Then the integrated imagery was fed to CNN to classify building function, building age, and land use. Cao et al. (2018) used FCN-VGG16 to fuse overhead imagery with SVI. Two individual FCN-VGG16 channels were set up for SVI and RSI, i.e., FCN_{SVI} and FCN_{RSI}. FCN_{RSI} aimed to extract the features from RSI, and FCN_{SVI} aimed to extract the features from SVI. The extracted features from FCN_{SVI} were fused with extracted features of FCN_{RSI} at the third convolutional block, and then finally fused imagery was fed to deconvolutional block for the final prediction map and obtained an overall accuracy of 78.10%.

2.3.5. Accuracy Assessment

In RS investigations, the output classification map accuracy is compared to reference data obtained from municipal datasets or data collected in the field or data delineated by RS-experts. Generally, the comparison has been made based on the kappa coefficient, overall accuracy, recall, precision, F1 score, and Jaccard Index (also known as Intersection over Union (IoU)) (Rahman & Wang, 2016) to evaluate output classification map statistical significance. In the case of slum mapping, there is one major problem in assessing the accuracy of the output classification map using slum reference data because different institutions define slum differently, resulting in the generation of different slum maps for the same area as discussed in Section 2.1. These uncertainties in the slum reference map negatively affect the classification accuracy of different slum mapping approaches. Therefore many researchers define their own definition of slums according to the local context of the study area and generate the slum reference map manually using image interpretation technique to assess the accuracy of the output slum classification map from different RS-based approaches (Kohli, 2015).

There are a variety of accuracy metrics have been used in previous slum mapping studies. This section discusses the different accuracy indicators used in different slum mapping studies during recent years. Table 2.1 shows the list of accuracy indicators that have been used in previous slum mapping studies adopted from Gao (2020).

Recent Studies	Approaches	Accuracy Indicators
Stark et al. (2019)	Deep learning (FCN)	IoU
Wurm et al. (2019)	Deep learning (FCN)	PPV, IoU
Persello & Stein (2017)	Deep learning (FCN)	OA, Recall
Verma et al. (2019)	Typical CNN	OA, Kappa estimate, IoU
Leonita et al. (2018)	Machine learning (SVM, RF)	OA, Kappa estimate, F1-Score
Kohli et al. (2013)	Grey-Level Co-occurrence Matrix	OA, Precision, Recall

Table 2.1: List of accuracy indicators used in previous slum mapping research adopted from Gao (2020)

3. STUDY AREA

This chapter provides the background of Jakarta. Section 3.1 describes the demographic scene in Jakarta, followed by explaining the slum dynamics under Section 3.2. Section 3.3 discusses the challenges in slum monitoring due to various terminology of slums and the evolution of the urban village. Finally, section 3.4 explains the selection of study area in Jakarta.

3.1. Introduction to Jakarta

Indonesia's capital city Jakarta is the second largest urban agglomeration globally (Martinez & Masron, 2020), including five regions and one regency. The population of Jakarta is 10.56 million in 2019, which means it was increased by 10% from the 2010 census (Martinez & Masron, 2020). Table 3.1 shows the region-wise population density of Jakarta.

Region	**Area in Km^2**	**Population**	**Population density in Km^2**
Central Jakarta	48,13	1.056.896	21.959
Western Jakarta	129,54	2.434.511	18.794
Eastern Jakarta	188,03	3.037.139	16.152
Southern Jakarta	141,27	2.226.812	15.763
Northern Jakarta	146,66	1.778.981	12.130
Thousand Islands	8,70	27.749	3.190
Total	662,33	10.562.088	15.947

Table 3.1: Region-wise population density of Jakarta
Retrieved from https://jakarta.bps.go.id/

3.2. Slum Dynamics in Jakarta

Jakarta is a rapidly urbanizing city and one of Indonesia's largest densely populated provinces (Martinez & Masron, 2020). It is Indonesia's economic center that attracts people from other parts of the country to search for work opportunities. Thus, an increasing population causes the scarcity of affordable housing inside the city that forces people to live in low-quality housing (Pratomo, Kuffer, Martinez, & Kohli, 2017), resulting in informal settlements' growth. Around 60% of Jakarta's population lives in informal settlements called kampungs (Pratomo et al., 2017). Figure 3.1 shows the location map of Jakarta.

Since 1997, Jakarta's government has been monitoring slum dynamics through ground-level surveys to update slum areas on the map. The measurement has been done using ten indicators: building material, population density, building density, building orientation, air circulation, clean water, sanitation, drainage, wastewater disposal, and type of roads (Pratomo et al., 2017). The result of ground-level surveys has been categorized into four slum classes (Figure 3.2): very light slums, light slums, medium slums, and heavy slums, but the criteria used by local government officials for dividing slums into different categories are not clear. Between 2014 – 2015, slum areas are reduced under the local government's policies in which slum areas were largely relocated, resulting in the drastic change in slum dynamics in Jakarta (Pratomo et al., 2017).

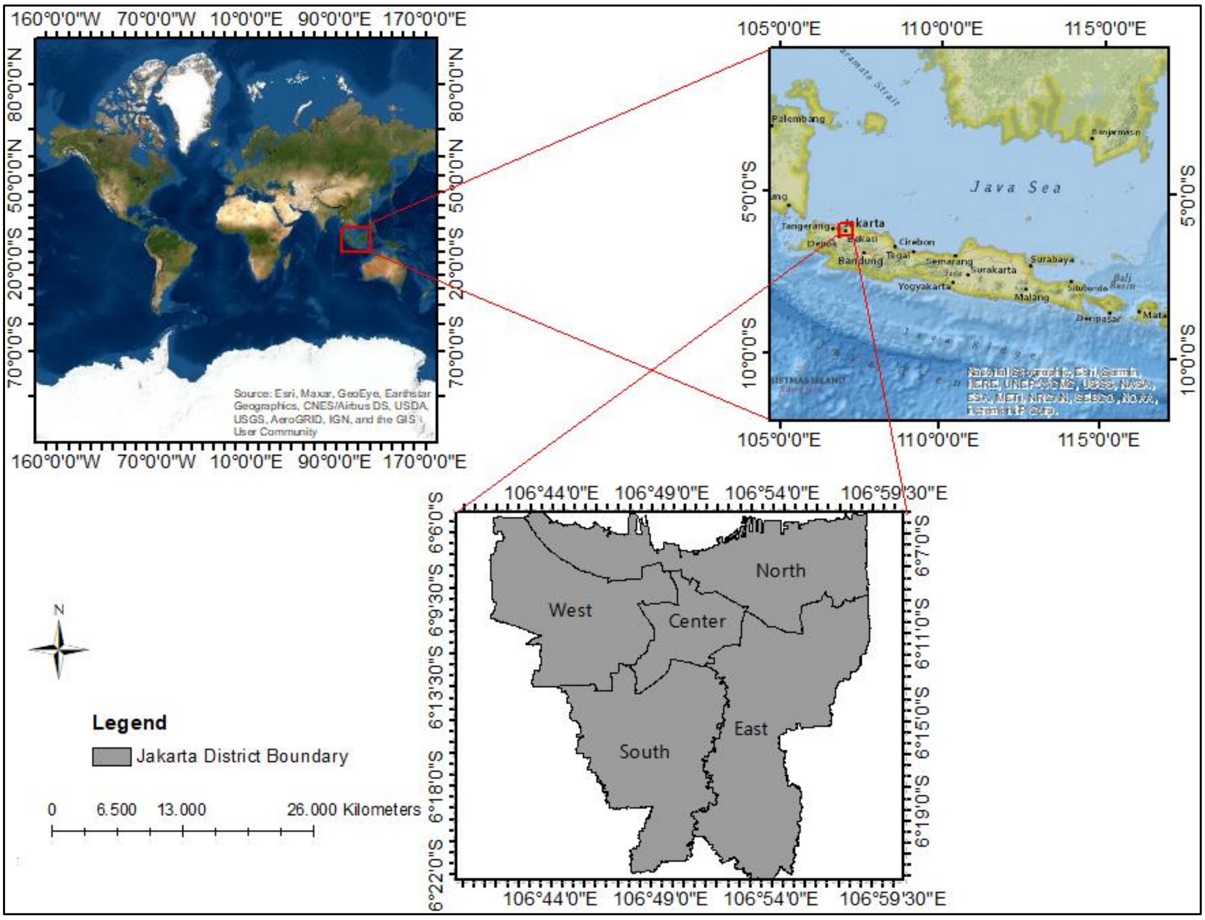

160°0'0"W 70°0'0"W 10°0'0"E 90°0'0"E 170°0'0"E
80°0'0"N
50°0'0"N
20°0'0"S
70°0'0"S
Source: Esri, Maxar, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community
105°0'0"E 110°0'0"E 115°0'0"E
5°0'0"S
10°0'0"S
Palembang
Banjarmasin
Java Sea
Tangerang
Jakarta
Depok
Bekasi
Bandung
Tegal
Cirebon
Semarang
Surakarta
Surabaya
Yogyakarta
Malang
Denpasar
Bali
National Geographic, Esri, Garmin, HERE, UNEP-WCMC, USGS, NASA, ESA, METI, NRCAN, GEBCO, NOAA, increment P Corp.
106°44'0"E 106°49'0"E 106°54'0"E 106°59'30"E
6°6'0"S
6°9'30"S
6°13'30"S
6°18'0"S
6°22'0"S
6°7'0"S
6°11'0"S
6°15'0"S
6°19'0"S
North
West
Center
East
South
N
Legend
Jakarta District Boundary
0 6.500 13.000 26.000 Kilometers

Figure 3.1: Location map of Jakarta

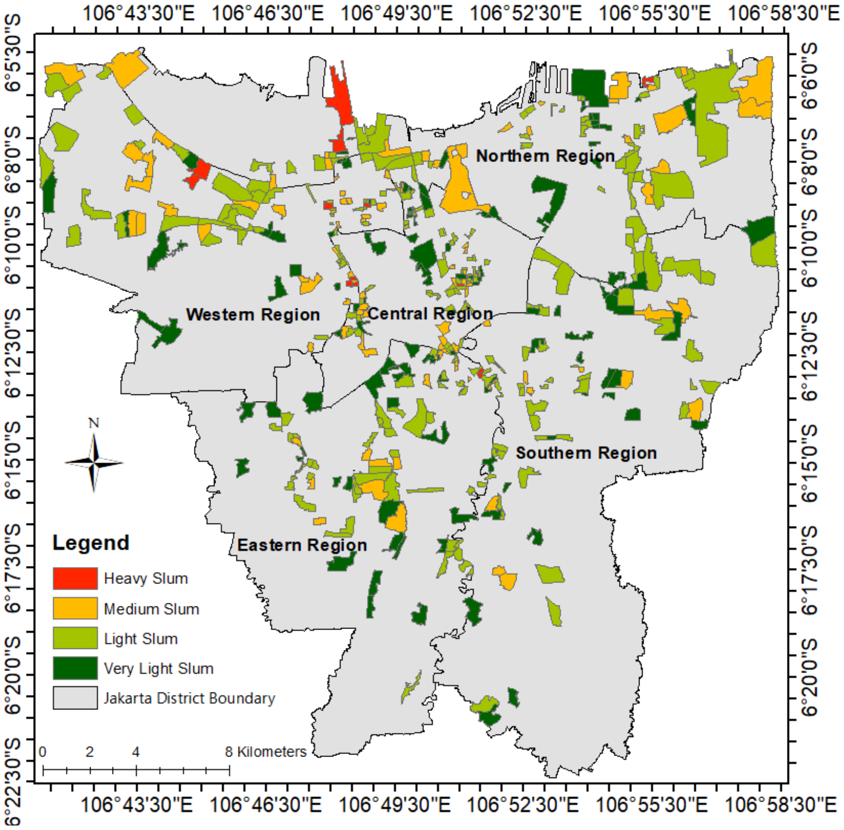


Figure 3.2: Official slum reference data of 2017 with different categories of slums, i.e., heavy, medium, light, and very light slums

3.3. Problems in Mapping Slum Dynamics

There are mainly two major issues to map slum dynamics in Jakarta. First, the different existing definitions of slums. Second is the presence of kampungs, also known as urban villages, where it is challenging to differentiate slum and non-slum areas.

3.3.1. Different Definitions of Slums

Indonesia is committed to aligning its development target with the 2030 global agenda of SDGs. Accordingly, the national government has included global agendas in the development planning policy and programs such as the National Medium-Term Development Plan (RPJMN) and its related budget (Minister of National Development Planning [MNDP] Indonesia, 2019). For example, one of the leading global agendas incorporated into the national development plans is "Improving the quality of housing and settlements" under RPMJN 2020 – 2024. However, the main emphasis is given to Goal 6- "Clean Water and Sanitation" (MNDP Indonesia, 2019), under which 100-0-100 (100% access to clean water, zero slums, and 100% access to sanitation) policy was implemented. The key to ensuring the success of this initiative is to keep track of slum reduction progress. However, different institutions in Indonesia at the national, regional, and local levels define slums differently.

The definition of slums is not universal, which makes it challenging to monitor slum dynamics. According to the Indonesian National Law on Housing No 1/2011, slums are divided into slum housing and slum neighborhoods (Irawaty, 2018). Slum housing is defined as an inadequate living space, whereas a slum neighborhood is defined as housing without basic amenities (Pratomo et al., 2017). Accordingly to national law, different institutions defined several indicators to measure slums (Pratomo et al., 2017). For example, the ministry of public works and public housing defines six indicators: coverage and quality of the road network, poor water quality, poor wastewater disposal, density and quality settlements, and the area size of inundation.

In contrast, Indonesia's central board of statistics defines four indicators: insufficient living space, poor quality of building materials, lack of access to drinking water, and poor sanitation. Likewise, Jakarta's local government also defines eight indicators: building layout and orientation, inadequate living space, quality of housing, garbage collection, sanitation, building density, unpaved/light roads, and air and light ventilation to measure slums based on national law definition. Thus using the different sets of indicators will lead to different measurements of slums. Indonesian national planning agency came up with a concept of legal and illegal slums, i.e., if the owner owns the house and the government recognizes it, then it is called a legal slum, whereas if the owner doesn't own the house and the government doesn't recognize it then it is an illegal slum. This research focuses on illegal slums. The comparison between the different institution definitions of slums is shown in Table 3.2

We have defined our definition of slums for this research based on some of the important indicators used by different governmental institutions and academic literature to avoid different interpretations from different conceptualizations of slums. The indicators used to define slums in this research are the absence of tenure security, temporary building materials, a dense area with lesser roads, unplanned layout, unpaved/light roads, building footprint area less than 60-meter square (m^2), poor roofing materials, near to industrial and warehouse area, proximity to rivers, railroads, swamps, and shrine, and less open green spaces, this will be discussed in detail in Section 5.2.

3.3.2. Presence of Kampungs (evolution of Kampungs)

Discussing the problem of mapping slums is closely related to the kampungs in Jakarta. About half a century ago, after the end of the colonial period, Jakarta faced rapid urbanization (Putranto, 2009). At that time, the government does not govern any planning institution (Rukmana, 2008), promoting irregular housing development. For example, many vacant land and agricultural fields turned into settlements, and some of the settlements are dominated by low-income groups called kampungs. More people moved towards Jakarta due to the rapid urbanization resulting in new kampungs or expanding the existing ones. The increasing population has put enormous pressure on the housing sector, and many people have to opt for substandard housing due to the local government's incapacity. This gradual growth made kampungs bigger and more heterogeneous with the middle-class income group (Pratomo et al., 2017).

Kampungs can be categorized into two types (1) legal kampungs have been provided the land rights and basic amenities although high-density characteristic doesn't change (Putranto, 2009). On the other hand, (2) illegal kampungs don't have land rights and basic amenities generally located along the railway line, riverbank, green paths and park, canal, and often in flood-prone areas (United Nations-Habitat [UN-Habitat], 2013), and these illegal kampungs are generally invisible from the city plans. Figure 3.3 and Figure 3.4 show the different types of kampungs, i.e., legal and illegal. However, legal and illegal kampungs may share some characteristics, i.e., high-density housing, making it quite challenging to categorize legal kampungs (non-slums) and illegal kampungs (slums).

Criteria	International[1]	National Law[2]	National Institution[3]	Susenas[4]	Local Institution[5]
Lack of Basic Amenities	Included	Included	Included	Included	Included
Lack Quality of Housing	Included	Included	Included	Included	Included
Inadequate Living Space	Included	Included	Included	Included	Included
Insecurity of Tenure	Included	-	-	-	-
Non-conformity with Spatial Plan	-	Included	-	-	-
Poor Socio-economic Condition	-	Included	-	-	-
Poor Accessibility	-	-	Included	-	Included
Hazardous Area	-	-	Included	-	-
Other	-	-	-	-	Included

Table 3.2: Comparison between the different institution definitions of slums for better understanding
Partially adapted from Pratomo et al. (2018)

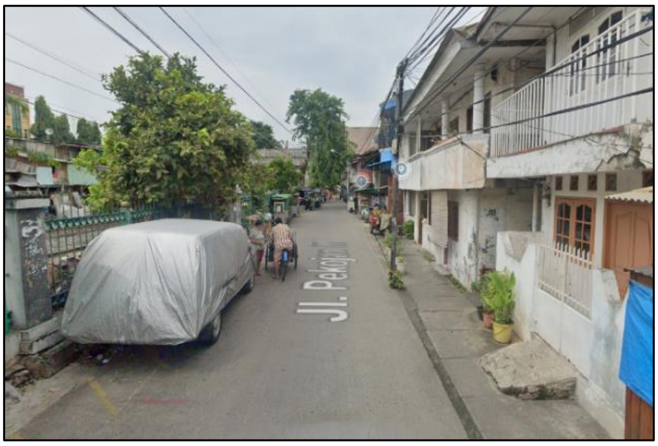


Figure 3.3: Legal Kampung in which owner owns the land rights on which they live, i.e., non-slum
Retrieved from Google street view images

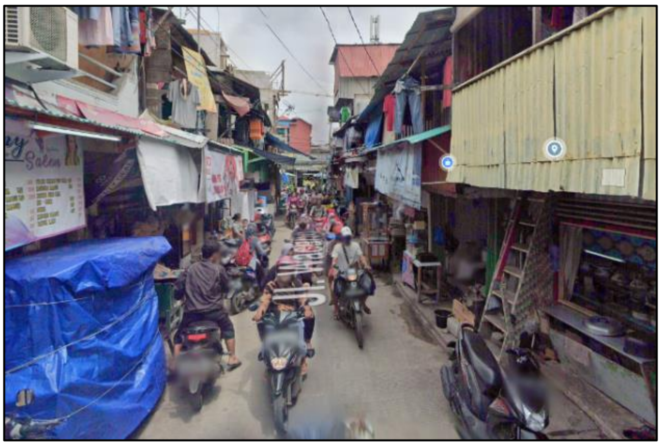


Figure 3.4: Illegal Kampung in which owner doesn't own the land rights on which they live, i.e., slum
Retrieved from Google street view images

[1] United Nations Habitat from UN-Habitat, 2018; UNDESA, 2017
[2] Government of The Republic of Indonesia from The World Bank, 2016
[3] Ministry of Public Works and Public Housing from Pangeran & Akbar, 2020
[4] Indonesian Central Board of Statistics from Central Bureau of Statistics, 2019
[5] Department of Building and Settlements DKI from Central Bureau of Statistics DKI Jakarta, 2020

3.3.3. Selection of Subset in Jakarta

The idea behind selecting the study area is to cover the major location of slums in Jakarta according to the official slum reference map of 2017. The western region has been selected as the subset of Jakarta for this research because it is the second-most densely populated region after the central region, as shown in Table 3.1. Central Jakarta is not chosen due to the Central Business District (CBD) because the presence of a large number of commercial places around the CBD with fewer residential and open spaces implies the probability of dense slums (heavy slum) will be minimum as shown in Figure 3.2. Therefore, the west region is chosen over the central region. Due to data availability constraints, we couldn't acquire data for the whole western region. Consequently, we modified the study area with a local expert's help (Mr. Jati Pratomo - Ph.D. Candidate, PGM Department, Faculty ITC, University of Twente) and acquired half of the west region with some part of the north region with heavy slums and a small part of the central region along the creek where the probability of slum will be maximum, as shown in Figure 3.5.

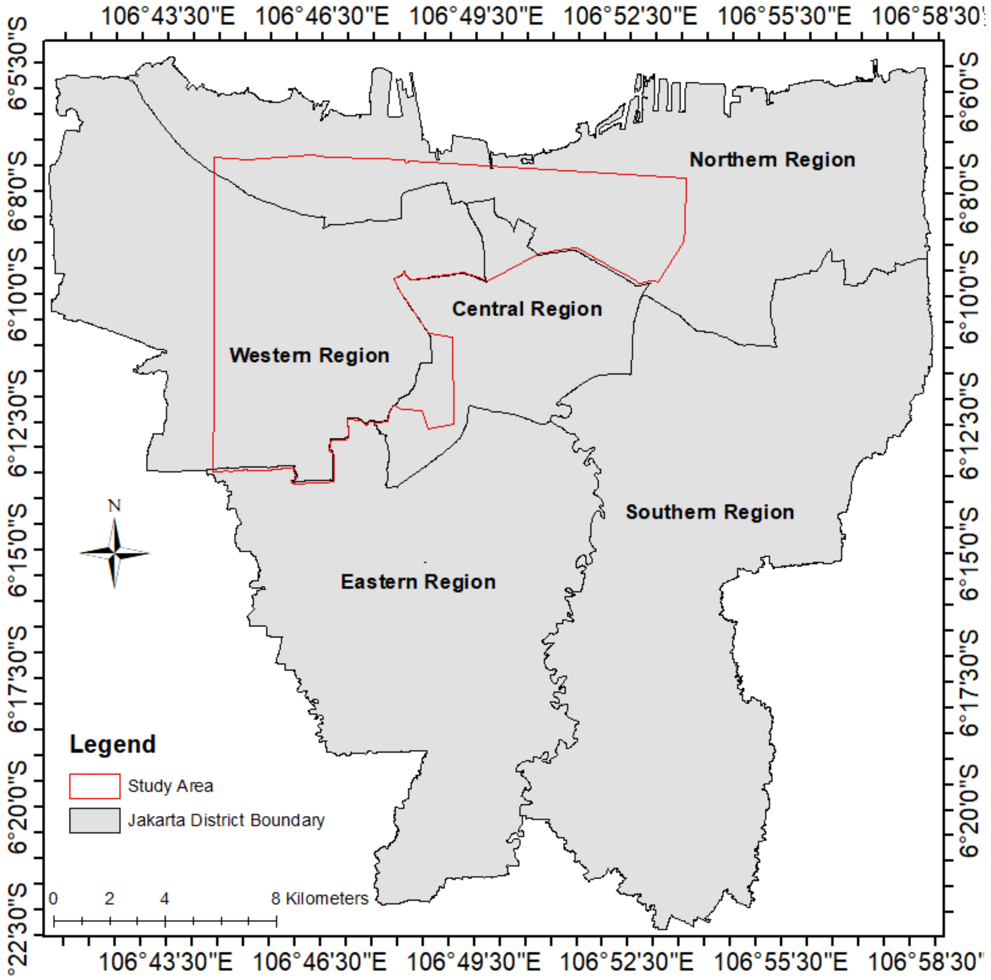


Figure 3.5: Extent of the study area and satellite imagery is highlighted with red, covering half of the west, some part of the north, and a small part of the central region (the satellite image was ordered according to the study area)

4. DATA

The chapter describes six different datasets used in this research. Section 4.1 describes the acquired VHR satellite imagery. Section 4.2 provides a brief explanation of the official reference data of 2017, and Sections 4.3, 4.4, and 4.5 discuss ancillary data such as road network, building footprints, and zoning data of Jakarta. Finally, section 4.6 describes the procurement of the Google Street View (GSV) imagery.

4.1. Satellite Imagery

This research concentrates on delineating slums in complex urban scenarios using VHR satellite imagery. We chose to explore the WorldView/GeoEye mission for VHR imagery. The imagery was procured from the European Space Agency (ESA) free of cost by sending a detailed project proposal (European Space Agency, 2021).

ESA approved the GeoEye mission satellite imagery of 125 km2. We tried to order the image in the same year (2017) as when the official slum reference data was prepared, but due to the data availability constraint, we have to shift the ordering date to the year 2018. The ordered image has a spatial resolution of 1.65 m for multispectral image with four bands (red, green, blue, and near-infrared band) and 0.41 m panchromatic image. ESA provided a pan-sharped image with a spatial resolution of 0.4 m. While procuring satellite imagery in tropical countries, the biggest issue is avoiding cloud interference in the ordered satellite imagery. Therefore we have chosen an option of cloud cover of less than 10%. The procured image specifications are shown in Table 4.1, and the extent of the ordered image and the procured satellite imagery are shown in Figure 3.5, and Figure 4.1shows the procured satellite imagery.

Image specification	**Description**
Bands	1 panchromatic and 4 multispectral (Red, Green, Blue, Near Infrared)
Resolution	0.4 m
Cloud cover	0%
Orthorectified on a scale	1:12000
Date of acquisition	March 2nd, 2018

Table 4.1: Detailed specifications of procured GeoEye satellite imagery from ESA

4.2. Slum Reference Data

In 2017 Jakarta's local government published official slum boundaries *(RW Kumuh)* based on different indicators mentioned in Section 3.3.1. Further, they have categorized slums into four different categories of slums such as heavy *(Berat)*, medium *(Sedang)*, light *(Ringan)*, and very light *(Sangat Ringan)*, as shown in Figure 3.2. Examples of GSV imagery of these different categories of slums are shown in Figure 4.2. The official slum map of 2017 was procured from Mr. Jati Pratomo. In addition, the official reference map was further tweaked for our analysis according to the definition of slum used for this research, as will be discussed in Section 5.3.1.1.

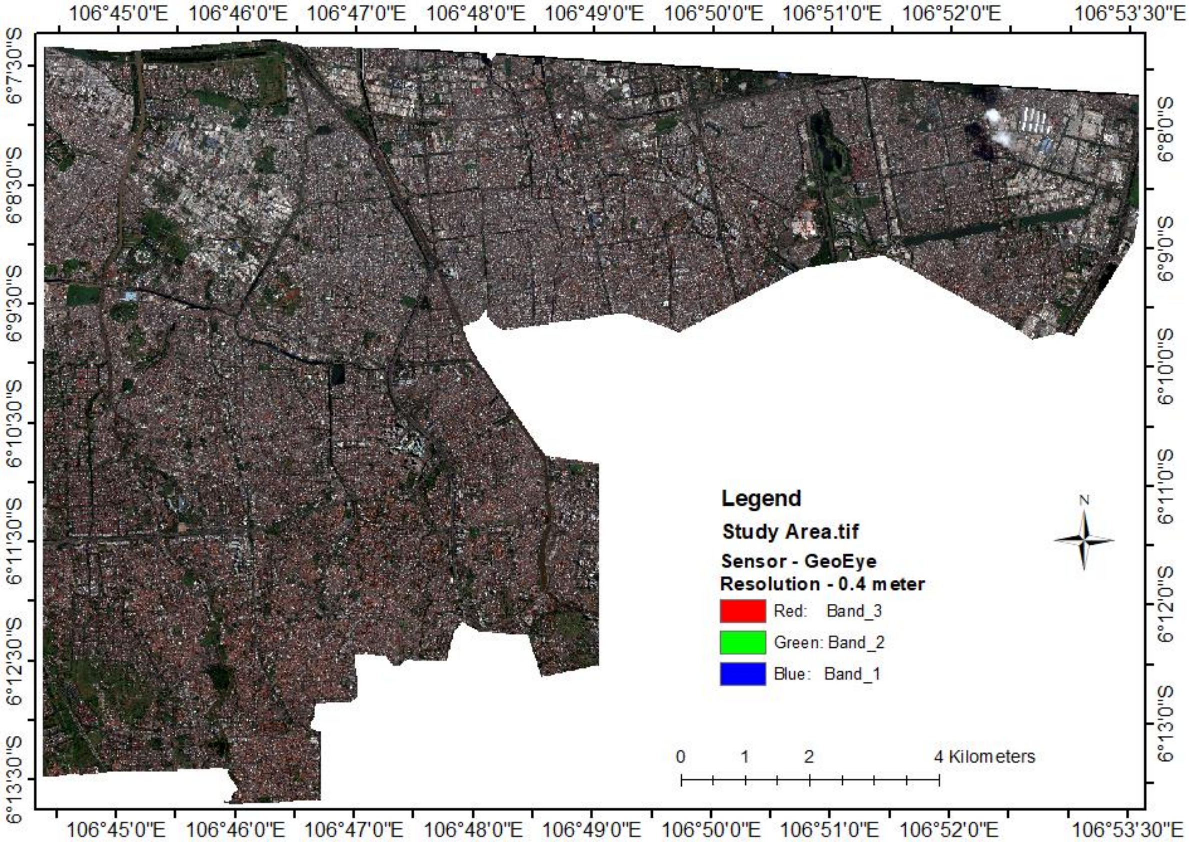
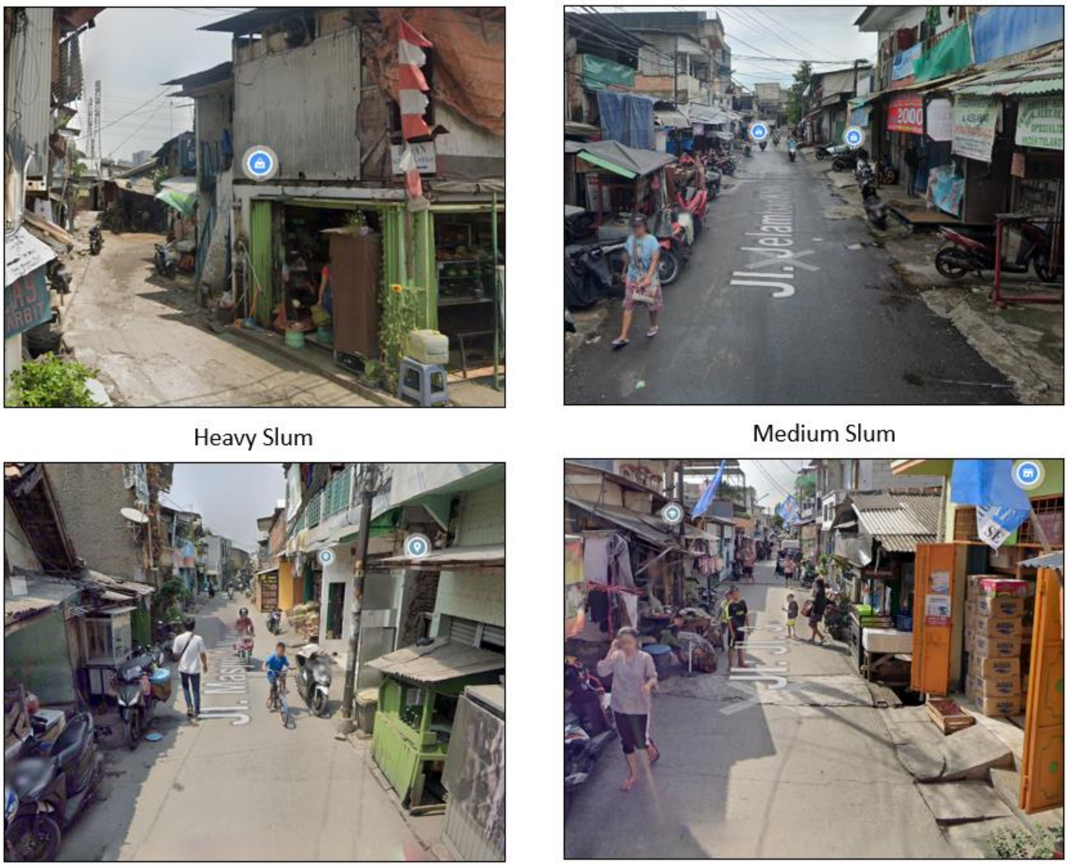


Figure 4.1: GeoEye satellite imagery of spatial resolution of 0.4 m procured from ESA



Heavy Slum

Medium Slum

Light Slum

Very Light Slum

Figure 4.2: Google street view imagery of different categories of slums according to official slum reference map of 2017

4.3. Road Network Data

The road network shapefile of Jakarta was downloaded from the official website of Jakarta's government (https://jakartasatu.jakarta.go.id/portal/apps/sites/#popup). Then the study area was clipped from the downloaded road network shapefile. The attribute table of the road network shapefile consists of road class *(kelas jala)* and information *(keterangan)*. The road classes are categorized into different classes such as major road *(artery and collector)*, minor road *(local)*, branch *(environment)*, and toll *(tol)* roads from which minor road and branch road are used for selecting the street view locations.

The information column consists of some important information related to road class, such as which roads are bridges, dirt roads (unpaved roads), links, and busways. Thus the information column is used to sort out the unpaved roads from the road network for tweaking the slum reference layer because the unpaved road is one of the indicators of slums in this research.

4.4. Building Footprint Data

The building footprint shapefile of Jakarta was downloaded from the official website of Jakarta's government (https://jakartasatu.jakarta.go.id/portal/apps/sites/#popup). Then the study area was clipped from the downloaded building footprint shapefile. The attribute table of building footprint shapefile consists of the building footprint area (shape area). The building footprint area is used to sort out the building footprint area less than 60 m^2 from the total building footprint area for tweaking the slum reference layer because building footprint is one of the indicators of slums in this research.

The concept behind taking the building footprint area less than 60 m^2 was when we overlayed the building footprint layer on the official slum reference map of 2017, we observed that the building area less than 60 m^2 overlays with the maximum number of buildings lying inside the slum boundary of the 2017 official slum reference map, which is called slums according to the local government, as shown in Figure 4.3.

4.5. Zoning Data

The zoning data shapefile of Jakarta was downloaded from the official website of Jakarta's government (https://jakartasatu.jakarta.go.id/portal/apps/sites/#popup). Then the study area was clipped from the downloaded zoning data shapefile. The attribute table of zoning data shapefile consists of the zone *(zona)*. The zone is used to sort out different zoning classifications such as green belt *(zona jalur hijan)*, city park *(zona tamon kota/ lingkungan)*, waterway *(zona terbuka biru)*, cemetery *(zona permakaman)*, and industrial and warehouse area *(zona industri dan pergudangan)* for tweaking the slum reference layer because zoning is one of the indicators of slum in this research.

The zoning classifications such as green belt, city park, waterway, and cemetery are areas where any kind of construction is prohibited, which means that the building constructed in these areas does not have land ownership. Thus, there is a high probability of finding slums in those areas.

4.6. Google Street View Images

The latitude and longitude in meters were generated for each street view location using ArcGIS. Then using the coordinate point of each street view location, images were obtained in cardinal directions using Google Street View Static API. For reference, SVI for the cardinal directions at one location in western Jakarta, where slums are present according to the official slum reference map of 2017, are presented in Figure 4.4. The selection of point locations of GSV will be discussed in-detail in Section 5.3.2.2.

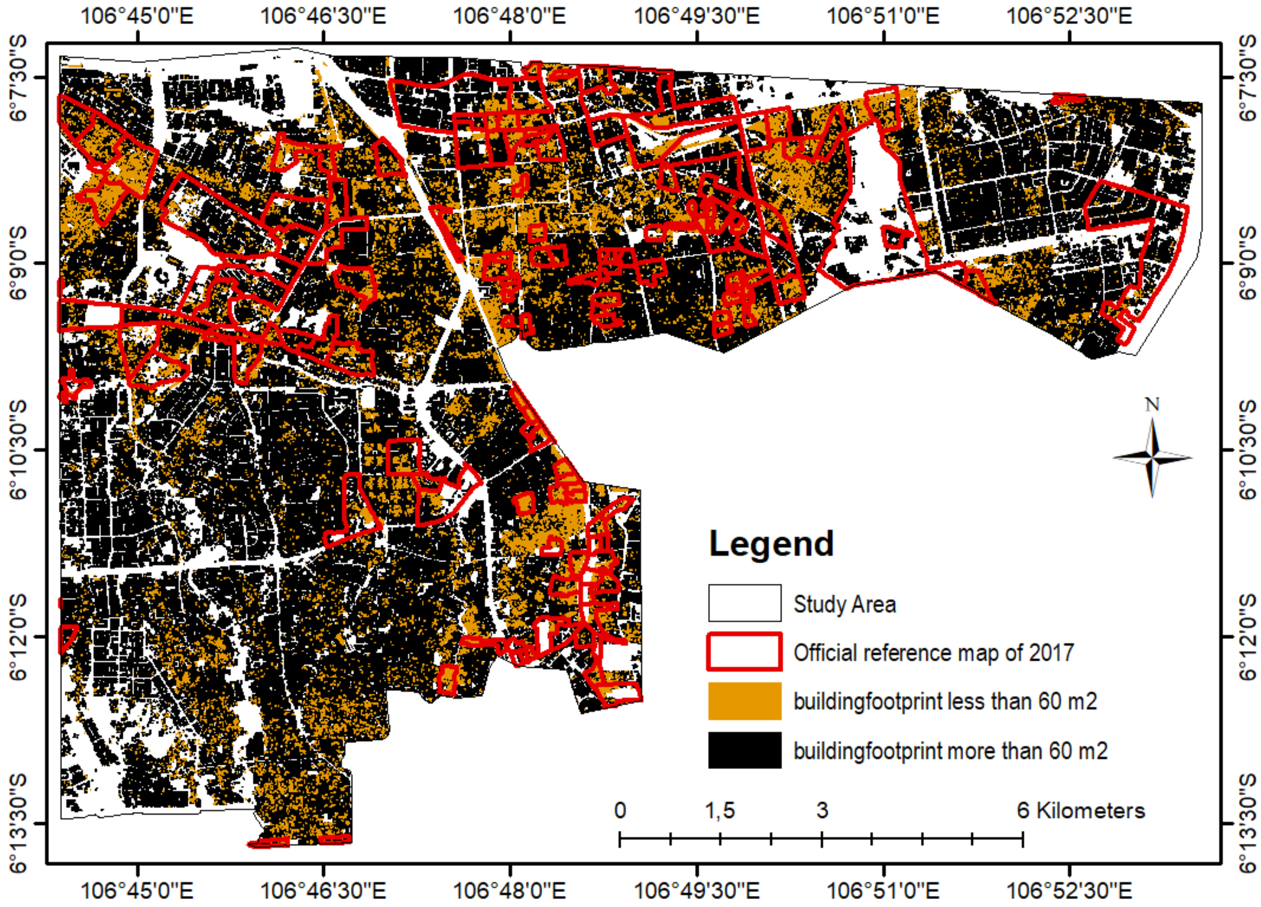


Figure 4.3: Official slum reference map of 2017 with building footprint area less than 60 m^2 marked in orange

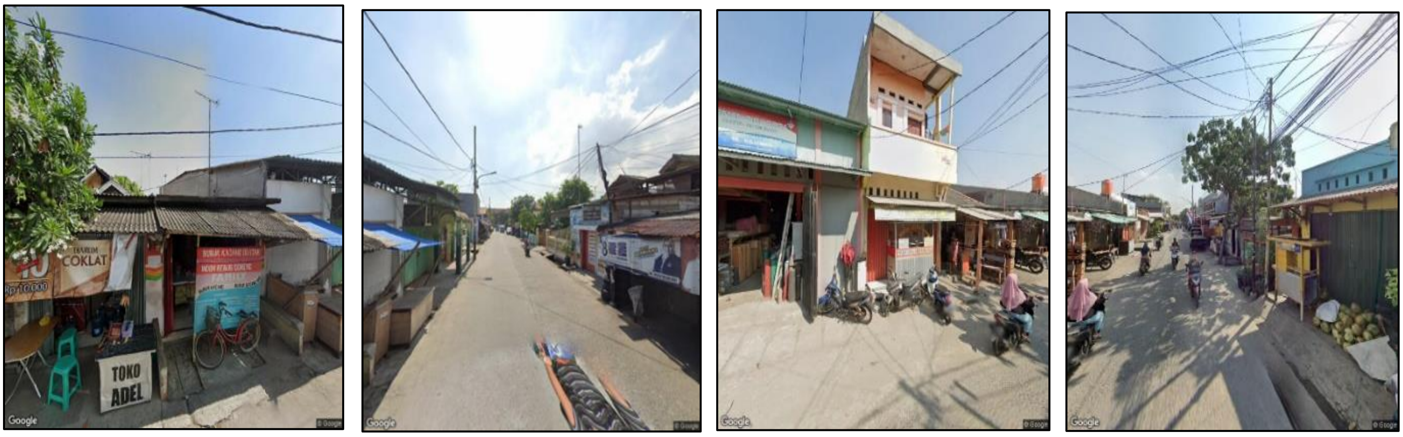


Figure 4.4: SVI images in the cardinal direction at one location in the study area were downloaded through Google API using latitude and longitude in meters

5. METHODOLOGY

This chapter discusses the methodology used to answer the research questions. Section 5.1 provides an overview of the research methodology. Section 5.2 explains the different steps taken for conceptualizing slums for this research. Section 5.3 explains the preparation of the dataset and execution of proposed architectures. Section 5.4 explains the selection of different accuracy metrics used to evaluate and compare the proposed architectures' results. Finally, section 5.5 provides the technical specification for executing proposed architectures

5.1. Overall Approach

This research is divided into three phases. In the first phase, we have identified the different characteristics of slums to conceptualize slums in our study area using RSI, SVI, and ancillary data. The second phase was divided into two steps: (i) pre-processing and (ii) experimental design. In the last phase, the accuracy of all models has been compared and analyzed. Figure 5.1 depicts the steps performed to attain the research objective.

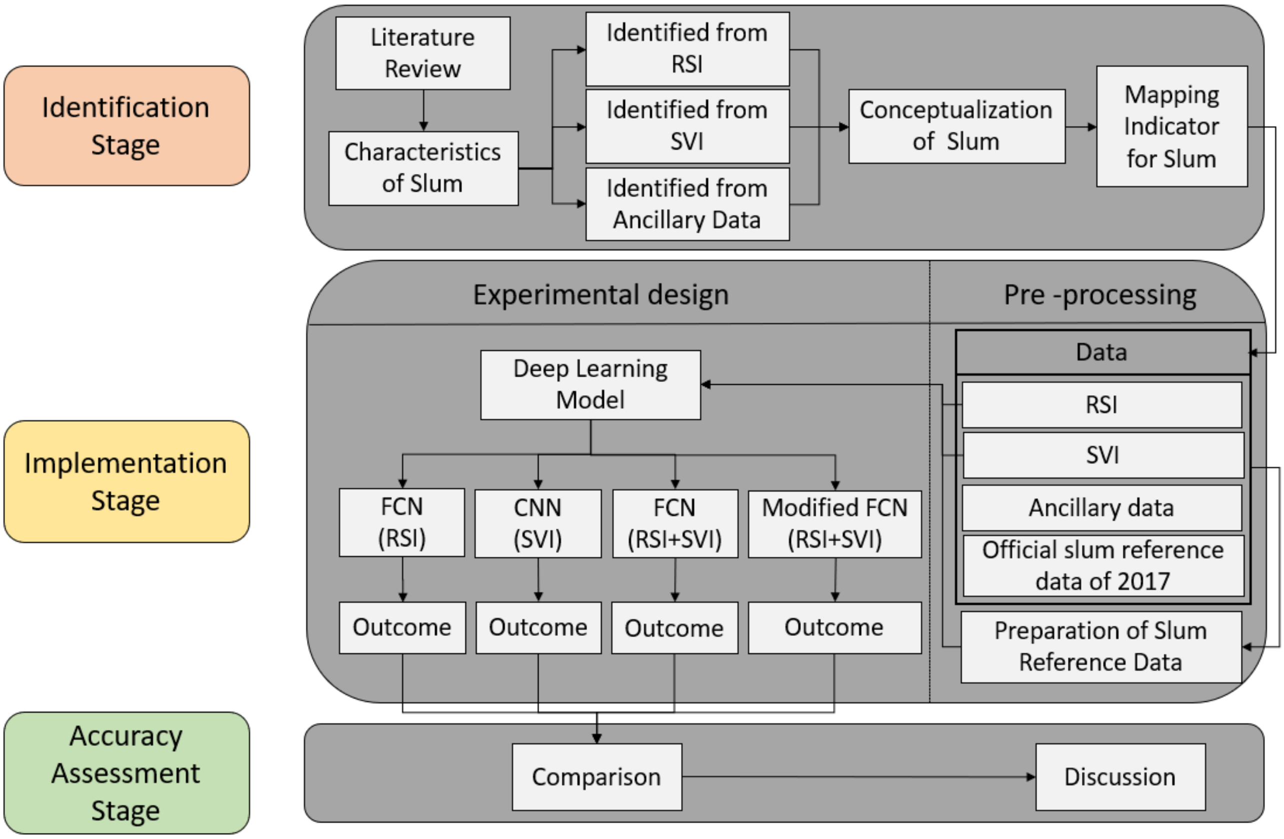


Figure 5.1: Overall approach of research (divided into three stages: identification, implementation, and accuracy assessment)

5.2. Identification Stage

This phase aims to determine the characteristics of slums in Jakarta, i.e., how slums can be seen in the real world and the image domain such as RSI and SVI with additional ancillary data. This stage is critical since the subsequent stages rely on the outcome of this procedure.

First, an extensive literature review was done to identify the characteristic of slums in two parts. In the first part, different governmental organization documents at global, national (Indonesia), and local (Jakarta) scales were reviewed, as mentioned in Section 3.3.1. We have gone through each set of indicators used by various governmental organizations to understand how they define slums. In the second part, different research papers have been reviewed to understand how they define slums in Jakarta's kampungs using different characteristics at the local level. Additionally, we discussed with a local expert (Mr. Jati Pratomo) to understand the slum characteristics in Jakarta at ground level. Based on the literature review findings, we have generated the list of indicators used in governmental documents and research papers to identify the characteristic of slums, as shown in Table 5.1. It is evident from Table 5.1 that physical characteristics play an important role in identifying slums.

Type of Characteristics	Characteristics	Adopted by	
Basic services	Lack of access to water and sanitation	International; National law; Susenas; Local Institution; Alzamil (2018); Zhu (2010)	🟧
	Lack of access to drinking water	International; Susenas	🟦
	Poor wastewater disposal	International ; National Institution	🟦
	Poor water quality	International ; National Institution	🟦
Socio-economic conditions	Socio-economic Conditions	National law	🟦
	Based on population	Legarias, Nurhasana, and Irwansyah (2020)	⬜
Others	Non-conformity with Spatial Plan	National law	🟦
	Absence of tenure security	International; Pratomo et al. (2017); Nurdiansyah (2018)	🟧
	Garbage collection	Local Institution	🟦
Physical	Lack of housing durability/ Poor quality of housing/ Poor quality of building materials/ Poor wall materials/ temporary building materials/ Poor roof materials	International; National law; National Institution; Susenas; Local Institution, Pratomo et al. (2017)	🟧
	Inadequate living space	International; National law; Susenas; Local Institution	🟦
	Coverage and quality of the road network / Dense area with lesser roads	National Institution	🟦
	Proximity to hazardous zone	International; National Institution	🟦
	High building density	Local Institution; Irawaty (2018); Legarias et al. (2020)	🟧
	Air and Light ventilation	Local Institution	🟦
	Unplanned layout/Irregular building	Local Institution; Pratomo et al. (2017); Alzamil (2018)	🟧
	Building orientation	Local Institution; Pratomo et al. (2017)	🟧

Type of Characteristics	Characteristics	Adopted by	
Physical	Unpaved/Light roads	Local Institution	🟦
	Small building size (building footprint)	Pratomo et al. (2017)	⬜
	Proximity to the river and railroads	Pratomo et al. (2017); Irawaty (2018); Alzamil (2018); Nurdiansyah (2018)	⬜
	Near to industrial and warehouse area	Pratomo et al. (2017)	⬜
	Hazardous location	Irawaty (2018)	⬜
	Less open and green spaces	Zhu (2010)	⬜

Government document and Research Paper	🟧
Government document	🟦
Research Paper	⬜

Table 5.1: List of characteristics of slum adopted by governmental documents at an international, national, and local level and research papers focusing on Jakarta

According to Pratomo et al. (2017), sometimes slums and non-slum areas in kampungs share the same physical characteristic, i.e., high-density housing, making it quite challenging to identify slums. Mapping slums in kampungs can be challenging with RSI alone because some of the physical characteristics that differentiate slums from non-slums cannot be captured through RSI, such as inferior building materials. Therefore to capture detailed physical characteristics, ground-level knowledge is needed. In this research, SVI is used for ground-level knowledge with RSI to identify the physical characteristics of slums in Jakarta's kampungs.

Only those slum characteristics were chosen from Table 5.1, which can be detected using RSI or SVI or available ancillary data such as road network data, building footprint data, and zoning data in our study area to conceptualize slums in this research. Table 5.2 shows the list of selected characteristics captured through RSI, SVI, and ancillary data, and Table 5.3 shows the translation of slum characteristics into the mapping indicators.

Type of characteristics	Characteristics	RSI	SVI	Ancillary data (Vector layer)		
				Road network	Building footprint	Zoning
Other	Absence of tenure security	No	No	No	No	Yes[6]
Physical	Temporary building materials	No	Yes	No	No	No
	Dense area with lesser roads	Yes	No	Yes[6]	No	No
	Unplanned layout	Yes	No	No	No	No
	Unpaved/Light roads	Yes	Yes	Yes[6]	No	No
	Small building size/building footprint	Yes	No	No	Yes[6]	No
	Poor roof materials	Yes	Yes (Partially)	No	No	No

[6] With the help of RSI

Type of characteristics	Characteristics	RSI	SVI	Ancillary data (Vector layer)		
				Road network	Building footprint	Zoning
Physical	Proximity to river, railroads, swamps, and shrines	Yes	Yes	No	No	No
	Near to industrial and warehouse area	Yes	No	No	No	Yes[6]
	Less open and green spaces	Yes	Yes (Partially)	No	No	No

Table 5.2: List of selected slum characteristics captured through RSI, SVI, and ancillary data in our study area for conceptualizing slums in the research

The selected characteristics shown in Table 5.2 for slum mapping are defined as follows:

Absence of tenure security: This indicator cannot be seen directly through the RSI and SVI. Therefore, tenure status was determined using zoning data to point out the illegal structures because Jakarta has strict zoning policies (Pratomo et al., 2017). In this research, zoning data (vector layer) was used to delineate illegal structures, i.e., slums, with the help of RSI. The attribute selected from the zoning data for finding the illegal encroachment was discussed in Section 4.5.

Temporary building materials: According to International; National law; National Institution; Susenas; Local Institution, slums in Jakarta's kampungs consist of temporary building materials such as iron sheets, wood blocks, plastic sheets, and low-quality construction materials. These temporary building materials are the ground-level characteristic that can be captured using the SVI. Therefore only SVI is used for identifying temporary building materials in our study area.

Dense area with lesser roads: According to National Institution, a high-density building area with less connectivity of roads is categorized as a slum. We have used RSI and road network data (vector layer) to delineate areas with less connectivity, .i.e, where the roads are not present compared to high-density buildings. The road network information was mainly inferred from road network data, according to which the area was delineated with the help of RSI, but in some areas where the roads were not present in the road network data, we have used visual interpretation for identifying the dense area with lesser road connectivity through RSI.

Unplanned layout: According to Local Institution; Pratomo et al. (2017); Irawaty (2018), slums in Jakarta consist of the unplanned/irregular shape of the building. The unplanned/irregular shape of the building was delineated through RSI, i.e., overhead imagery.

Unpaved/ light roads: This indicator was used by the Local Institution for delineating slums. This indicator can be observed through RSI with road network data, RSI individually, and SVI. In this research, we used RSI with road network data (vector layer) to find out which roads are unpaved in Jakarta according to which slums were delineated with the help of RSI. The attribute selected from the road network data for finding the unpaved roads was discussed in Section 4.3.

Small building size/building footprint: According to Pratomo et al. (2017), slums in Jakarta consist of the small size of the building footprint. The small building size can be seen through RSI individually and RSI with building footprint data. We have used building footprint data (vector layer) to quantify building footprints less than 60 m^2 and delineated slums with the help of RSI, as discussed in Section 4.4.

Poor roof materials: According to Pratomo et al. (2017), slums in Jakarta consist of inferior roofing materials such as iron sheets and asbestos sheets which were further confirmed through RSI and SVI, but we have also found plastic covering material used as roofing materials at some places in slums while exploring SVI. The slums were delineated on RSI with the help of visual interpretation through RSI and SVI.

Proximity to river, railroads, swamps, and shrines: According to Pratomo et al. (2017); Alzamil (2018); Irawaty (2018); Nurdiansyah (2018), many slums in Jakarta are present close to the river bank and rail lines which was further confirmed using RSI and SVI. We have added two more landmarks, such as swamps and shrines, because slums can be clearly located near them while exploring SVI. The slums were delineated on RSI with the help of visual interpretation through RSI and SVI.

Near to industrial and warehouse area: According to Pratomo et al. (2017), there is a high probability of unskill or low-skill workers' lives in slums (illegal kampungs) near industrial and warehouse areas. This indicator was observed through RSI individually and RSI with zoning data. In this research, we used zoning data (vector layer), i.e., industrial and warehouse areas, to delineate the nearby illegal kampung settlements with the help of RSI.

Less open and green spaces: According to Zhu (2010), slums in Jakarta's kampungs are densely packed structures overlapping each other, i.e., the probability of any green space or open space is lower in slum areas, which was confirmed through RSI and SVI. The SVI has limited coverage along the road, due to which this indicator was partially explored through SVI. The slums were delineated on RSI with the help of visual interpretation through RSI and SVI.

Characteristics	Mapping indicators
Absence of tenure security	• Ancillary data: Zoning data
Poor wall materials	• SVI: Iron sheets, Wood-blocks, Plastic sheets, and Low-quality construction materials
Dense area with lesser roads	• Ancillary data: Road network data • RSI Shape: Compactness
Unplanned layout	• RSI Shape: Compactness
Unpaved/Light roads	• Ancillary data: Road network data • RSI Shape: Compactness
Small building size/building footprint	• Ancillary data: Building footprint data • RSI Shape: Compactness
Poor roof materials	• RSI Tone: Iron sheets and Asbestos sheets • SVI: Iron, Asbestos, and Plastic sheets
Proximity to river, railroads, swamps, and shrines	• RSI Association: Proximity to River, Railroads, Swamps, and Shrines • SVI: Proximity to River, Railroads, Swamps, and Shrines
Near to industrial and warehouse area	• RSI Association: Near to the industrial and warehouse area • Ancillary data: Zoning data
Less open and green spaces	• RSI Association: less open and green spaces • SVI: less open and green spaces

Table 5.3: Translation of selected slum characteristics into the mapping indicators[7] for tweaking official slum reference map of 2017

[7] The characteristics in SVI can be seen through the naked human eye because the SVI's are true color composite images, and for interpreting RSI images, visual elements: tone, shape, size, association, geographical features are used.

5.3. Implementation Stage

The section starts with the pre-processing step in which the preparation of the relevant data is briefly explained for further analysis. Then the next step of experimental design is briefly described.

5.3.1. Pre-processing of Image

The pre-processing is divided into two steps. In the first step, the official slum reference map of 2017 was modified using the characteristics of slums discussed in Table 5.2. In the second step, the ground truth dataset for training and testing was prepared using the tweaked slum reference map generated in the previous step for further analysis.

5.3.1.1. Preparation of Tweaked Slum Reference Map

In the official slum reference map of 2017, some areas did not align with the definition of slums used in this research, as shown in Figure 5.2 and Figure 5.3. Figure 5.2 shows that some of the well-developed areas, such as large green spaces and high residential buildings, are classified as slums in the official slum reference map of 2017. Whereas in Figure 5.3, the industrial and warehouse areas are also classified as slums in the official slum reference map of 2017. Therefore, the official slum reference map was tweaked for this research according to our definition of slums using different mapping indicators as mentioned in Table 5.3: Translation of selected slum characteristics into the mapping indicatorsTable 5.3.

As shown in Figure 5.2, the official slum reference map has four categories of slums, but the tweaked slum reference map has an additional category "unknown" because when we started digitizing slums using different mapping indicators, as mentioned in Table 5.3, we found that some areas in the study area are not categorized as slums in the official slum reference map of 2017 but still possess the characteristics of slums according to our definition of slums, as shown in Figure 5.4. Therefore we have digitized those areas and categorized them as an unknown type of slum because we don't know the basis on which the local government of Jakarta differentiates slums into different categories, such as heavy slums, medium slums, light slums, and very light slums. Further, we have shared our generated slum map with the local expert (Mr. Jati Pratomo) to get an insight into the areas delineated as slum and especially the unknown categories of slums, and he confirmed that the slum layer is correctly classified according to his knowledge.

The tweaked slum map was used in this research to generate ground truth training and testing datasets for the DL models and finally used for finding the accuracy of the predicted slum maps. Figure 5.5 shows the tweaked slum reference map used in this research.

5.3.1.2. Preparation of Ground Truth Training and Testing Dataset

As we discussed above, the tweaked slum reference layer has five categories, but the primary objective of this research is the binary classification of the areas into slums and other (non-slum). Slums are represented as polygons in the tweaked slum reference shapefile. Therefore tweaked slum shapefile is converted to the raster into two classes (slum and other) with the exact spatial resolution of the RSI.

The data preparation for the training and testing dataset is based on previous studies of FCN and many trials. The slum raster was cut into 12 tiles with equal rows and columns, i.e., 2000 x 2000 pixels. The total area covered on the ground by each tile is 800 m x 800 m. The tiles are arranged manually to balance class distribution in each tile, i.e., we have tried to distribute the classes (slum and other) equally into each tile by manually adjusting them. Figure 5.6 depicts the arrangement of training and testing tiles according to tweaked slum reference data. Among the 12 generated ground truth tiles, 10 were used for training, and 2 were used for testing the FCN networks in the research.

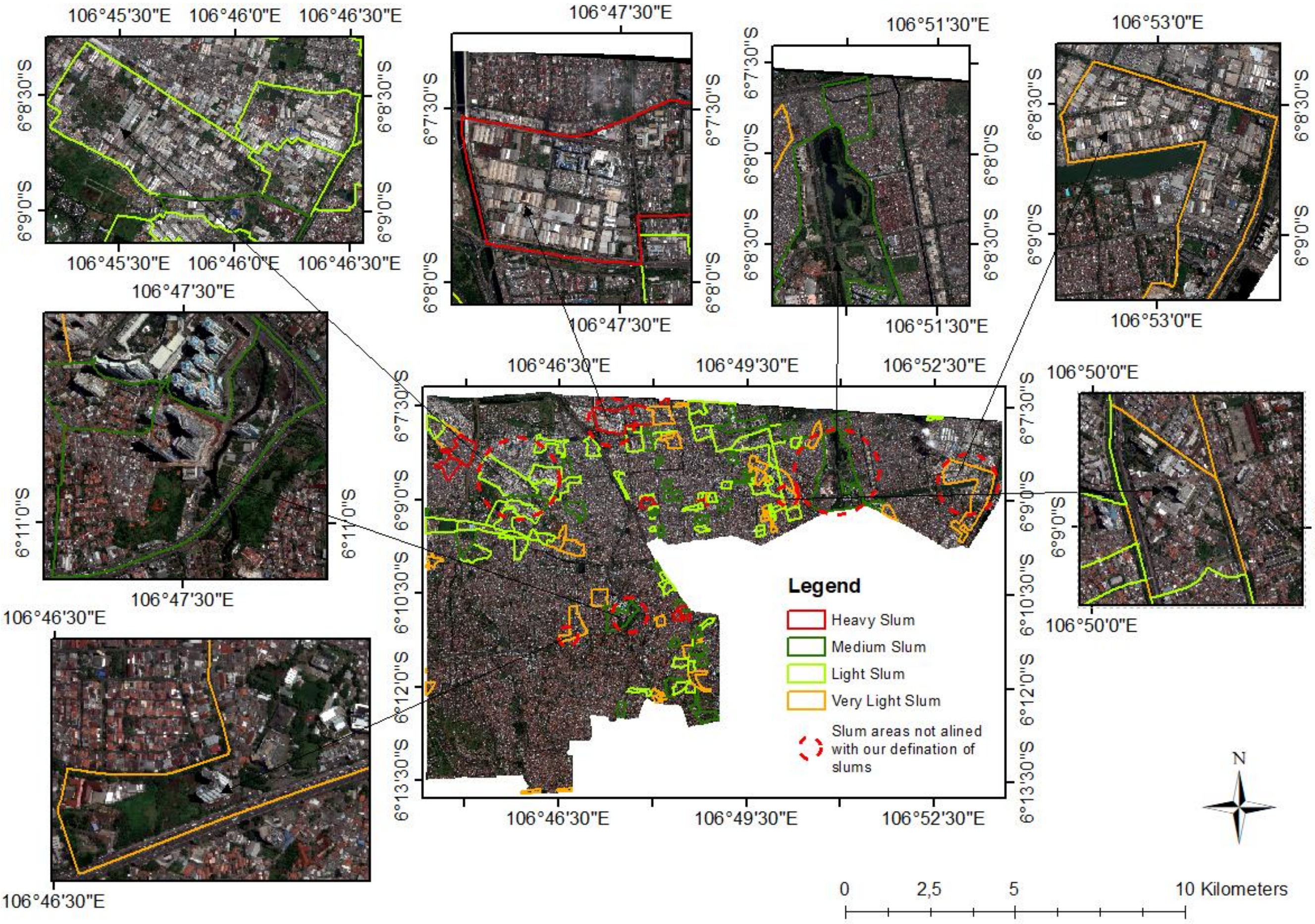


Figure 5.2: Contradicting areas such as large green spaces and high residential buildings are classified as slums in the official slum reference map of 2017

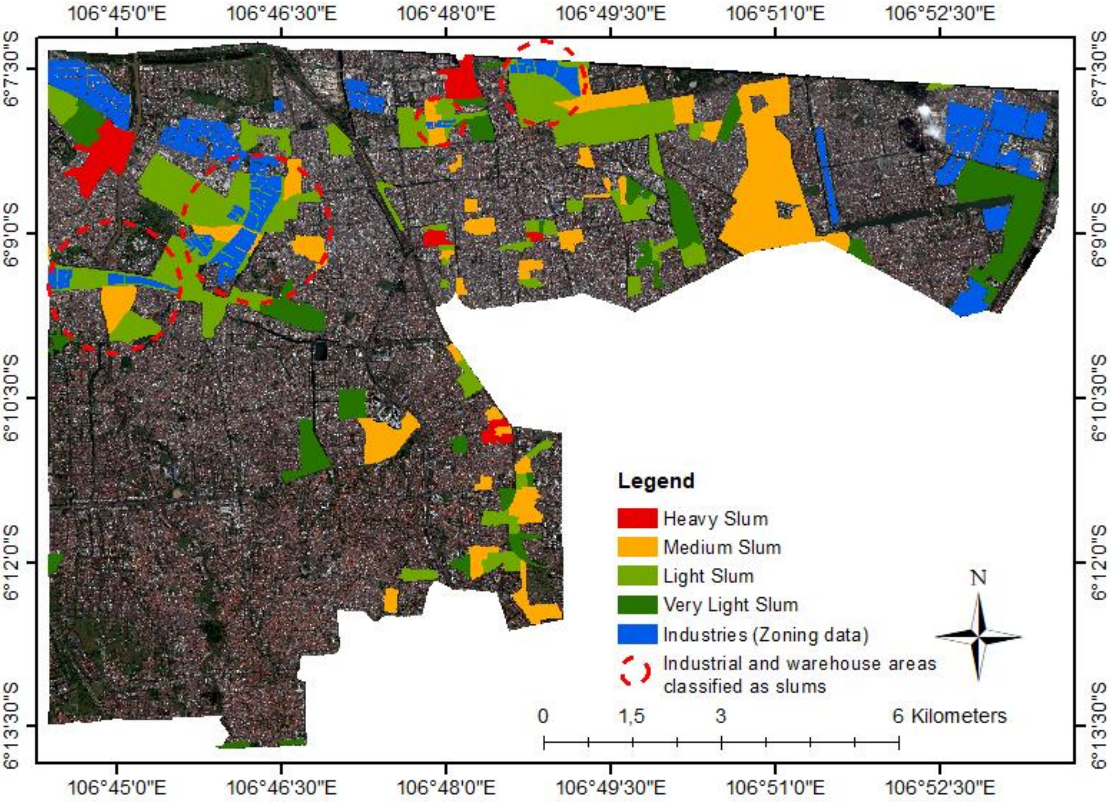


Figure 5.3: Industrial and warehouse areas are classified as slums in official slum reference map of 2017

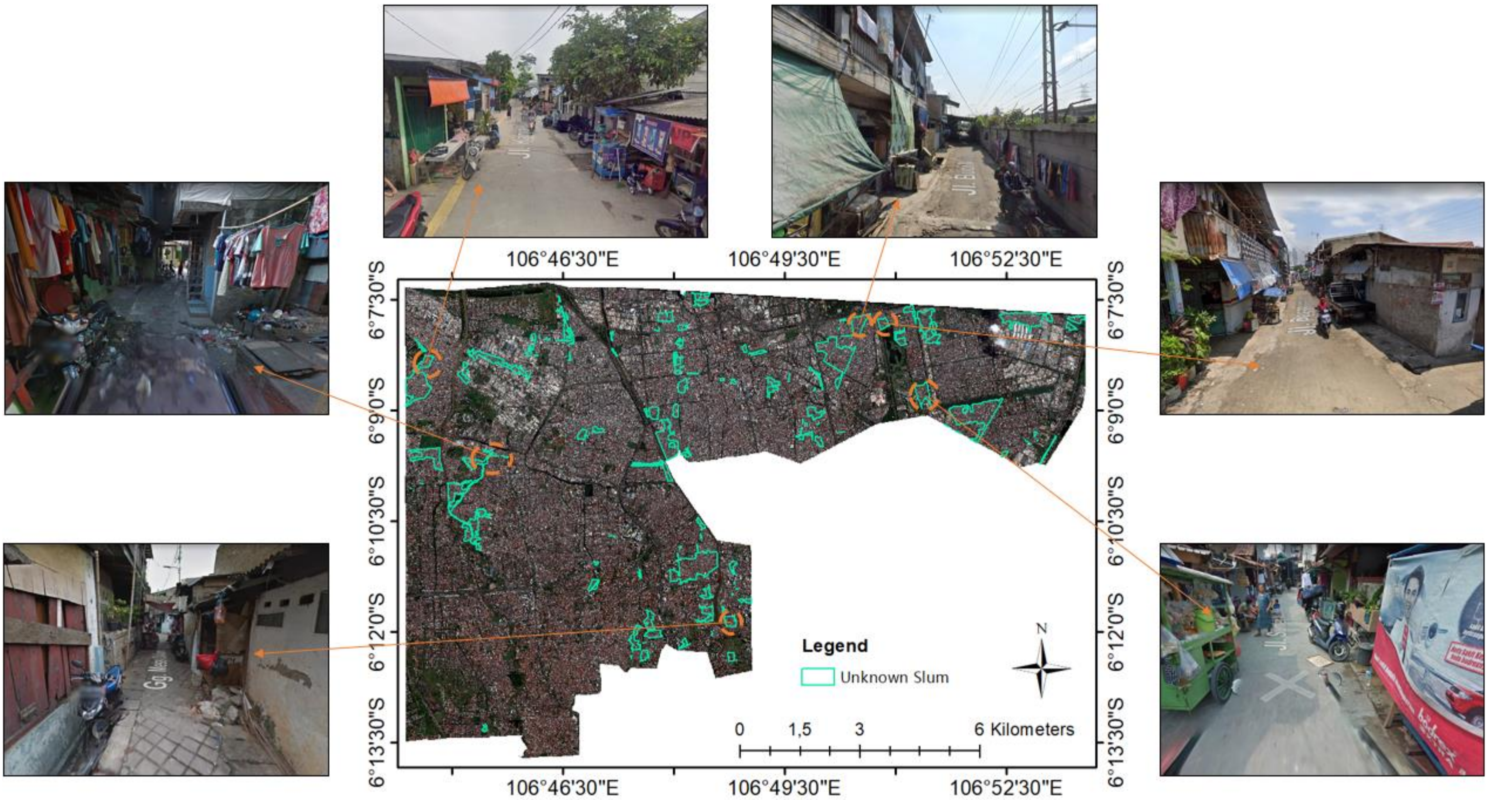


Figure 5.4: GSV images of unknown slum areas which are characterized as slums according to our definition of slums, but they are not delineated as slums in official slum reference map of 2017

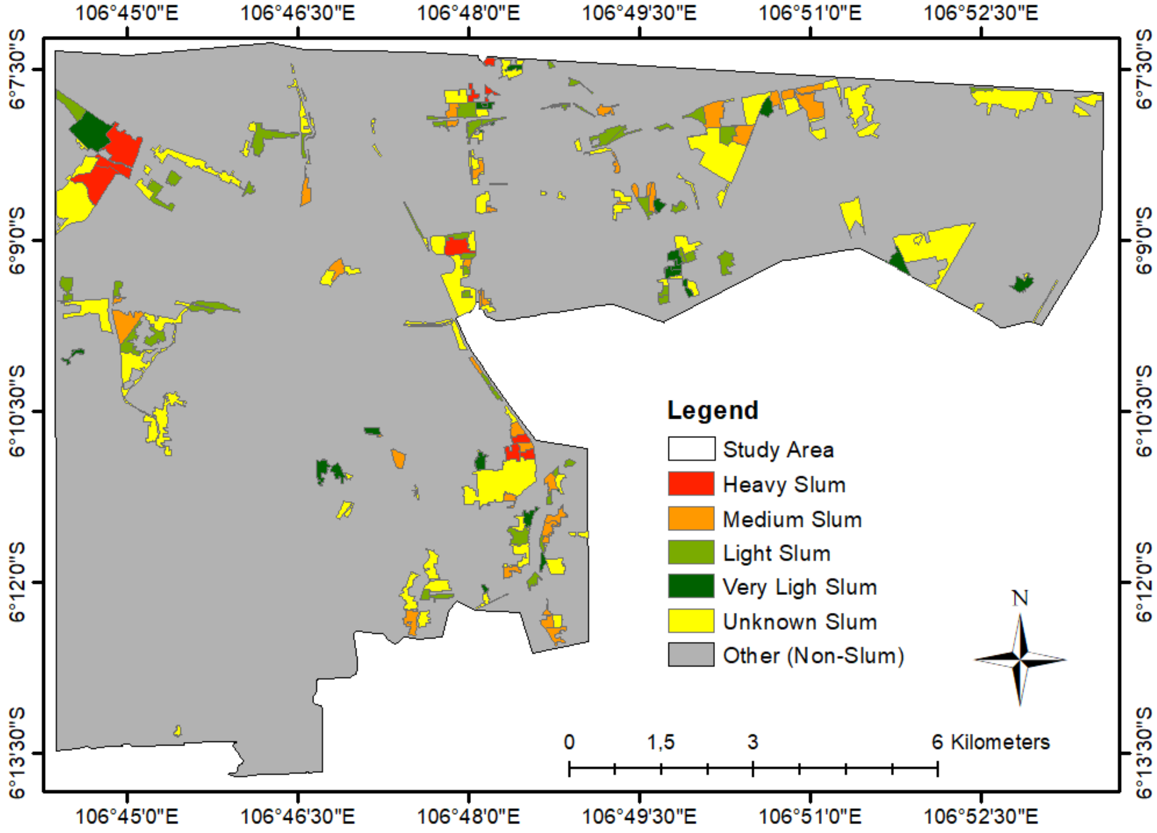


Figure 5.5: Tweaked slum reference map generated by using mapping indicators for this research

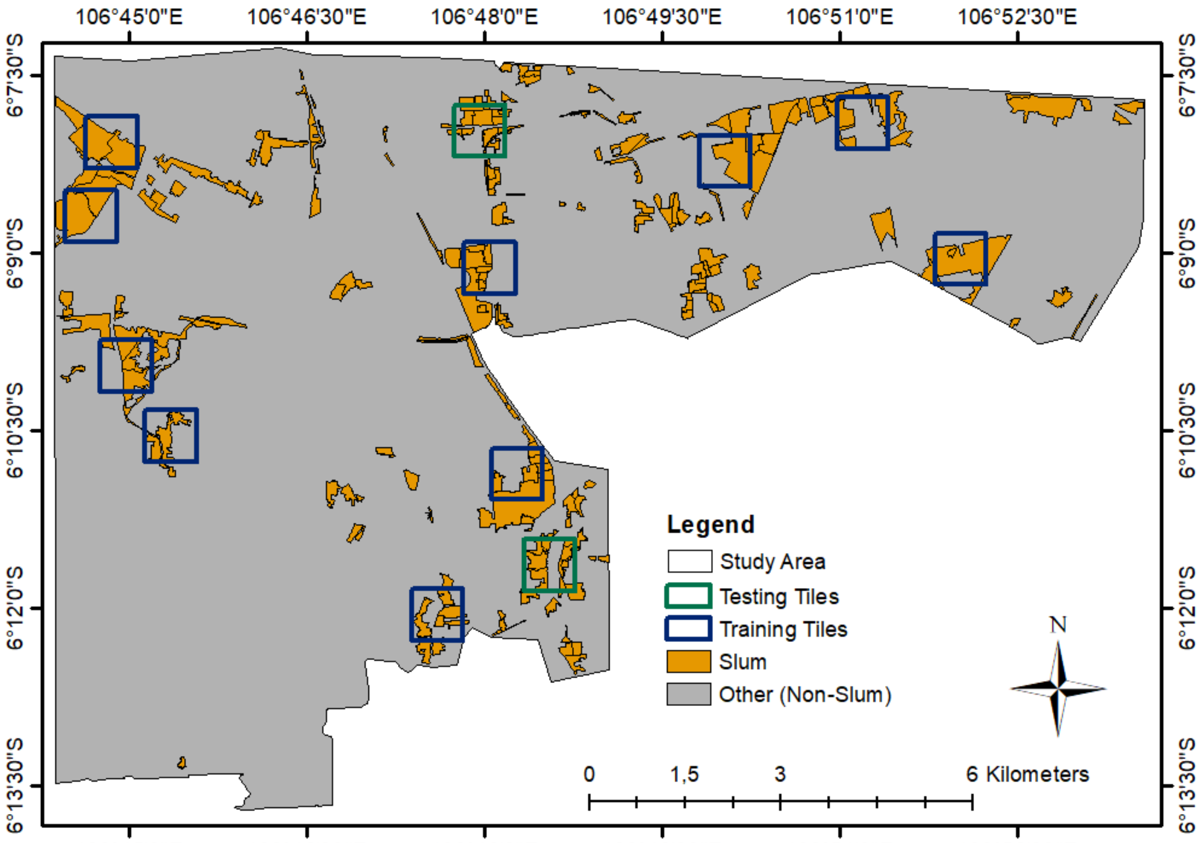


Figure 5.6: Arrangement of 12 ground truth tiles according to tweaked slum reference data in which 10 are used for training and 2 are used for testing FCN network

5.3.2. Experimental Setup

This section is divided into three major sections based on datasets: (i) RSI, (ii) SVI, and (iii) integration of RSI and SVI. Further, each section is divided into 3 sub-sections, i.e., network selection, data preparation, and training of the selected network.

5.3.2.1. Remote Sensing Imagery

I. FCN Architecture

Persello & Stein (2017) introduced FCN-DK for slum mapping in which FCN-DK6 outperforms all the other FCN-DKs architecture, as mentioned in Section 2.3.2. Furthermore, the FCN-DK can support the n-number of input bands for training, whereas FCN-VGG19 supports only three bands (Long et al., 2015). Therefore the FCN-DK6 architecture was used for this research. A detailed explanation of the proposed FCN-DK6 architecture is given in Annexure-I.

II. Data Preparation

The RSI was clipped according to 12 ground truth tiles generated based on the tweaked slum reference map, as discussed in Section 5.3.1.2. The snapshot of RSI and ground truth tiles prepared for training FCN-DK6 is shown in Figure 5.7.

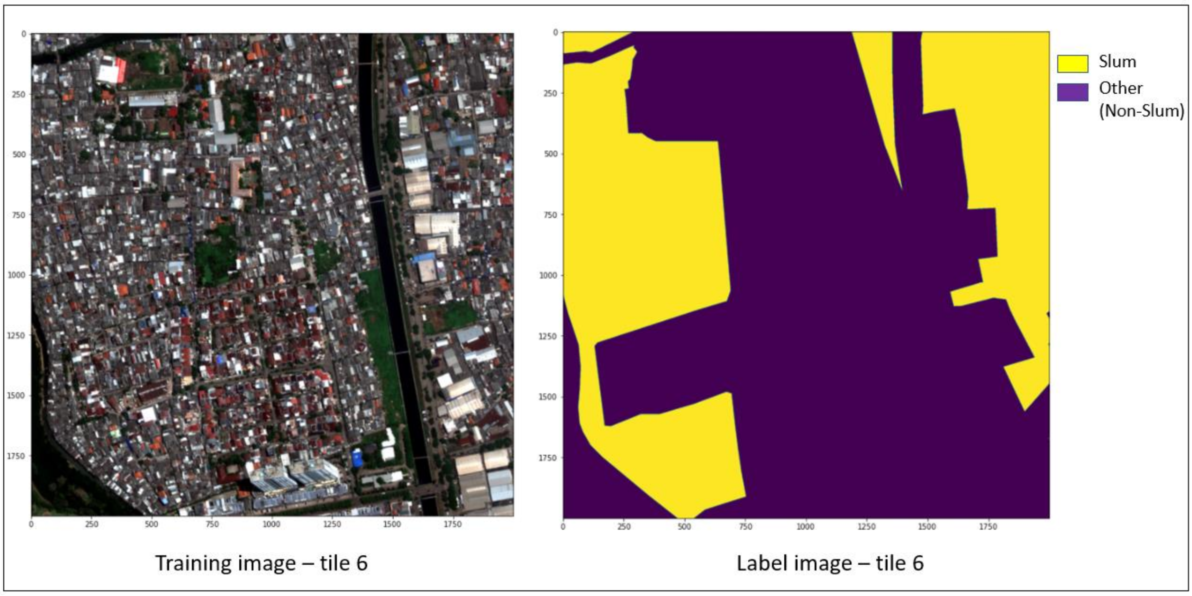


Figure 5.7: Snapshot of RSI and ground truth tiles prepared for training FCN-DK6

III. Training Network

The samples were systematically extracted from each training tile by splitting them into non-overlapping equal-area patches with the dimension of 125 x 125 pixels. Based on many trials, the size of the patch dimension was fixed. Thus 256 training patches have been created for each training tile. Finally, the network configuration shown in Table 5.4 was used to train FCN-DK6, and the weights are randomly initiated. We have also used EarlyStopping methods from Keras library to monitor validation accuracy with the patience of 10 epochs, i.e., the model will stop training if it doesn't see any rise in validation accuracy in the past 10 epochs.

Number of epochs	300
Batch size	64
Validation split	0.30
Optimizer	Stocastic Gradiant Desent (SDG) Learning rate: 1x 10^{-5} Momentum: 0.9

Table 5.4: Network configuration used for training FCN-DK6 in this research

5.3.2.2. Street View Imagery

I. CNN Architecture

As discussed in Sections 2.3.1 and 2.3.3, training CNN from scratch can be time-consuming, resource-intensive, and require the disposal of a vast database. Therefore, a pre-trained network on the Places365 dataset was fine-tuned in the context of Jakarta to recognize urban scenes from SVI.

The Places dataset was introduced in the Places project and it consists of 434 scenes representing 98% of different types of man-made and natural scenes that a person can encounter, such as park, lawn, arena, desert, forest, etc. (Zhou, Lapedriza, Khosla, Oliva, & Torralba, 2018). The Places dataset was composed of four steps: first, the images were downloaded through an online image search engine using WordNet synonyms set for each scene class. Second, the annotators were provided with a specific definition of each class and 500-1200 ground truth images per class as a reference to label the downloaded image database. Third, the images that were not classified manually in the previous step were classified using AlexNet (deep learning scene classifier) with a validation accuracy of 32%. The classified images with predicted class confidence of more than 0.8 were further annotated manually, and the rest of the images were dumped due to the lower predicted class confidence value. Fourth, the separation between similar classes was improved manually. Finally, the data set was finalized with 10 million labeled images with 434 categories. Only 365 categories were selected from 434 categories with 4000 images for each category to generate the Places365 dataset. We have selected the Places365 dataset because slum is included as one of the categories in the dataset. We could not get the specific definition of slum used in the Places365 dataset. In contrast, we have explored the sample images of slums in the Places365 dataset and found out the characteristics of slums represented in the images, such as inferior building materials, low-quality roads, high-density housing, hazardous locations, poor roofing material, as shown in Figure 5.8. The four subset datasets that were generated from the Places365 dataset are Places365-Standard, Places365-Challenge, Places205, and Places88, and we have used the Places365-Standard dataset in this research.

Zhou et al. (2018) trained different CNN architectures such as ALexNet, GoogLeNet, VGG16 with the Place365-Standard dataset. The training has been done on nearly 1 million images, 50 images per class for validation, and 900 images per class for testing. They finally concluded that the VGG16 performed better than others networks. Therefore the same VGG16 network with 5 convolutional blocks and 1 classification block was used for this research. From now on in this research, VGG16 will be addressed as Places365-VGG16 because VGG16 was pre-trained on the Places365-Standard dataset. A detailed explanation of the proposed Places365-VGG16 architecture is given in Annexure-I.

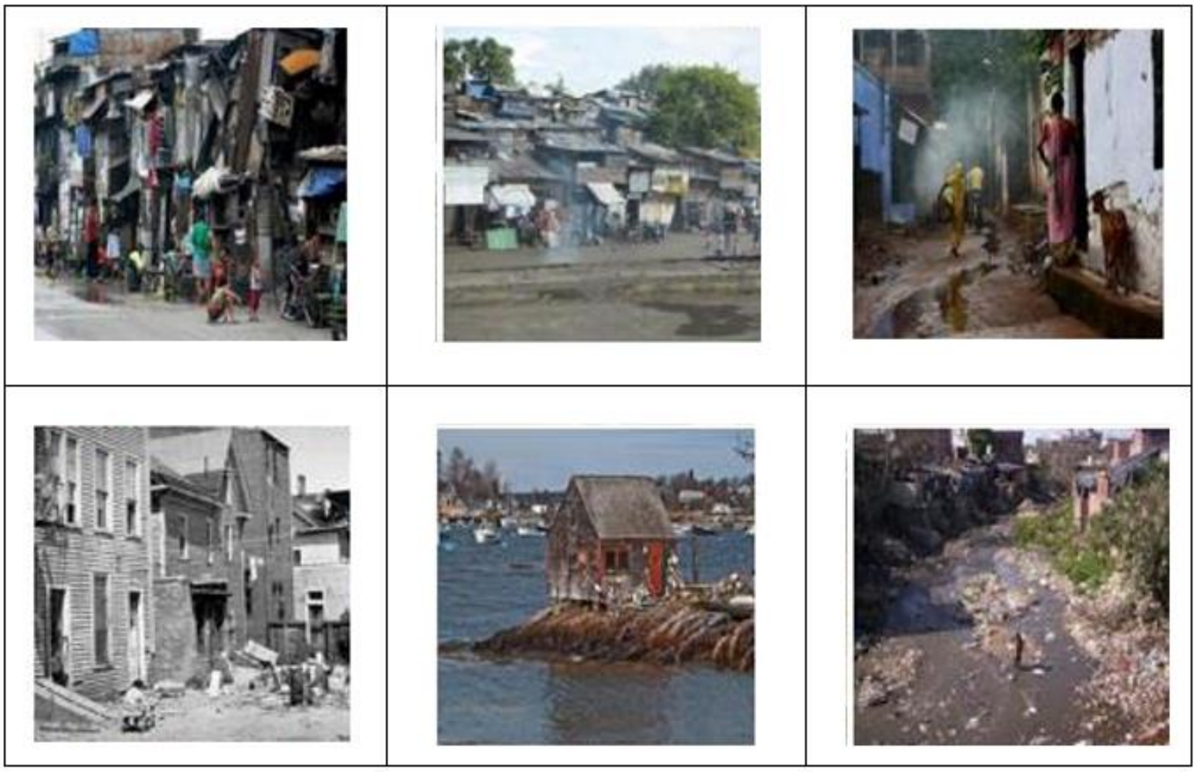


Figure 5.8: Sample images of slum category in Places365 dataset
(images were used for understanding the definition of slum used for generating slum categories in the Place365 dataset)

II. Data Preparation

As discussed in Section 4.3, minor road and branch road networks were used to generate random points on the road in the study area. As shown in Figure 5.5, slums cover a much smaller proportion of the city's surface area than non-slum (other). Therefore, we have generated random points in slum and non-slum areas individually and tried to generate an equal number of random points in both areas. We got one optimum solution through many trials in which the coverage of random points was distributed over slum and non-slum areas with an approximately equal number of random points using random points along line tool in QGIS. For slum areas, we have generated two random points per feature (road) with a minimum distance[8] of 25 m and a global minimum distance[8] (between the previously generated points on roads) of 10 m. For non-slum areas, we have generated one random point per feature (road) with a global minimum distance[8] of 50 m. The points on the boundary of slum and non-slum areas were removed from the analysis using spatial queries (select by location) in ArcGIS to ensure that the SVI capture at those points should only cover either slum or non-slum areas. For example, if the point lies on the boundary of slum or non-slum, then the SVI captured at that point will cover both slum and non-slum areas, which can generate incorrect results in further analysis. Finally, we obtain 6903 random points for slum areas and 7339 points for non-slum areas, as shown in Figure 5.9.

The randomly generated points for slum and non-slum are called Google Street View (GSV) Locations from now on in the thesis. Image in the four cardinal directions was obtained for each GSV location through the Google API. After fetching the images, we have got only 19796 images for slum and 24,596 for non-slum, and the rest were no data points, i.e., the GSV images are not present at those locations. The downloaded images were thoroughly checked for the anomalies such as completely dark images and images taken inside the buildings, and those images were deleted from the dataset. Finally, we have individually selected 19748 images for slum and non-slum, i.e., four images were taken at each GSV location. In total, 4937 GSV locations were selected for slum and non-slums individually. The whole image dataset is split into training, validation, and testing datasets, as shown in Figure 5.10.

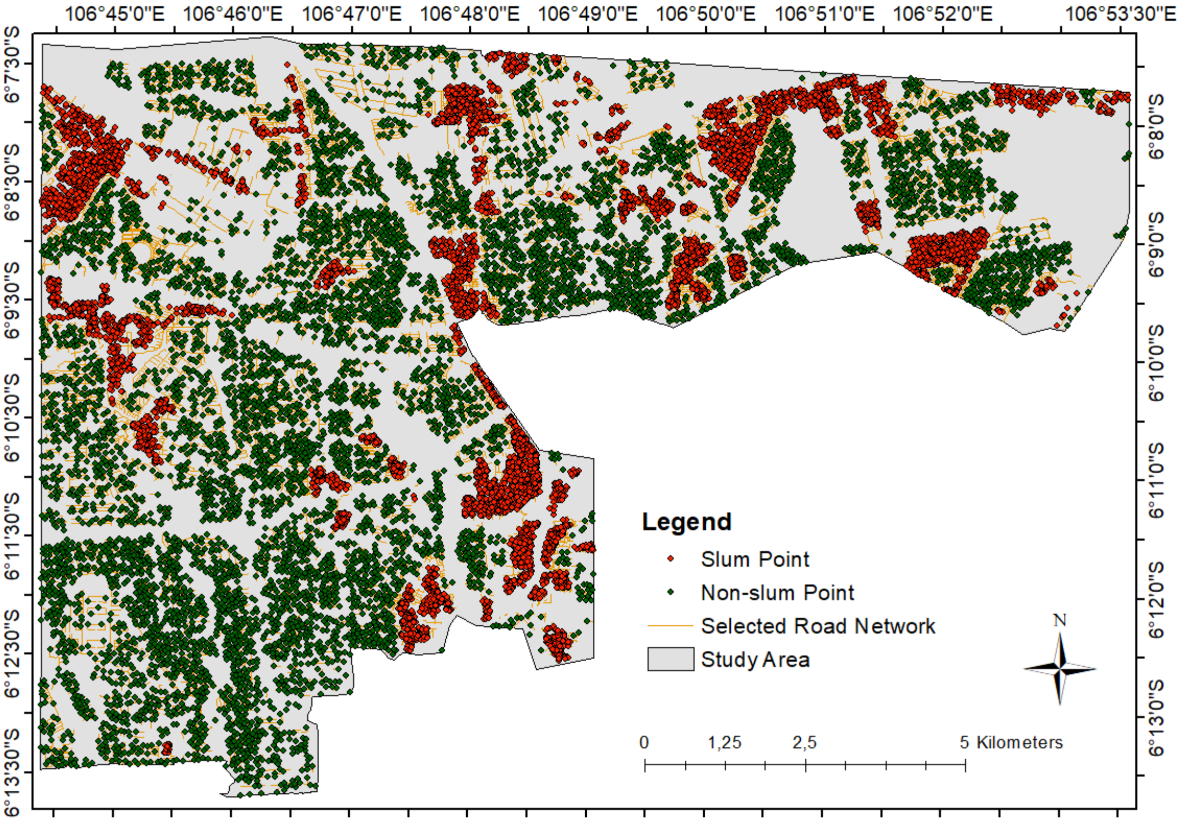


Figure 5.9: Distribution of random points on roads in slum and non-slum areas (green are non-slum point and red are slum point)

[8] with the minimum distance between points, only points on the same line feature are considered, while for the global minimum distance between points, all previously generated points are considered.

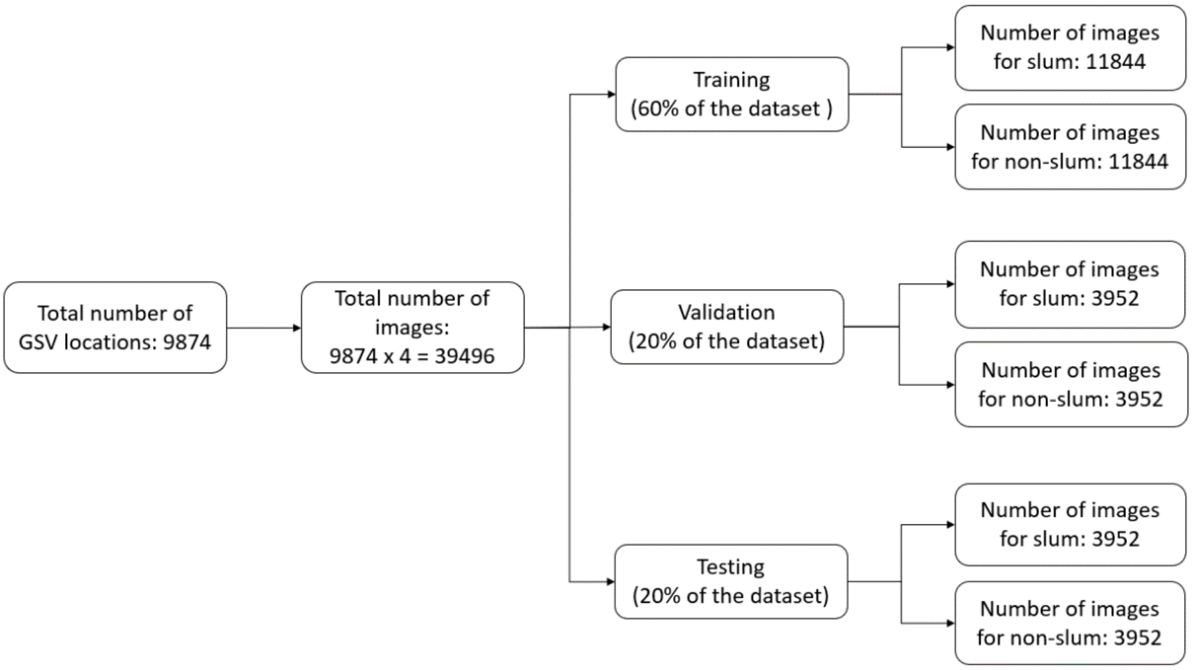

Total number of GSV locations: 9874
Total number of images: 9874 x 4 = 39496
Training (60% of the dataset)
Validation (20% of the dataset)
Testing (20% of the dataset)
Number of images for slum: 11844
Number of images for non-slum: 11844
Number of images for slum: 3952
Number of images for non-slum: 3952
Number of images for slum: 3952
Number of images for non-slum: 3952

Figure 5.10: Distribution of data (GSV images) for training, validation, and testing for Place365-VGG16

III. Training Network

Places365-VGG16 is a sequential model, i.e., all the layers are arranged sequentially in the model. First, the ImageData Generator was imported from the Keras library because it imports the data with their labels and is also used for data augmentation functions. Augmentation functions such as rescale, horizontal flip, and shear range are used for the training dataset, whereas only rescale augmentation function is used for validation and testing. The important thing about ImageData Generator is that it does not change the stored data; thus, it alters the data while passing it into the model. Slum and non-slum data are kept separately in different folders for training, validation, and testing.

To better understand slums in Jakarta, the Places365-VGG16 model was fine-tuned by training some layers and leaving other layers frozen. For this research, we have frozen the 5 convolutional blocks and fine-tuned the classification block. Finally, the Places365-VGG16 network configuration shown in Table 5.5 was used for fine-tuning Places365-VGG16, and the weights are initiated with Places365 weights. We have also used EarlyStopping methods from Keras library to monitor validation accuracy with the patience of 10 epochs, i.e., the model will stop training if it doesn't see any rise in validation accuracy in the past 10 epochs. Initially, the number of the epoch was set to 300 for training, but the model was terminated at the 49th epoch because the validation accuracy didn't increase in the past 10 epochs.

Number of epochs	300
Batch size	64
Optimizer	Stocastic Gradiant Desent (SDG) Learning rate: 1x 10^{-4} Momentum: 0.9

Table 5.5: Network configuration used for training Places365-VGG16 in this research

5.3.2.3. Integration of Remote Sensing and Street View Imagery

The feature maps from SVI were generated for integrating RSI with SVI. The generation of the feature maps consists of two main steps: (i) feature extraction and (ii) spatial interpolation.

- Feature extraction

Features from GSV images were extracted using fine-tuned Places365-VGG16 network to identify slum and non-slum areas. The features were extracted from the dense layer (dimensionality of 128) before the last fully connected layer of the fine-tuned Places365-VGG16 network; thus, 128 features were extracted for each GSV image. However, it is hard to study the 128 features for each GSV image. Therefore to reduce the feature dimensionality of the GSV image, the feature reduction technique is used. There are two main reasons for using the feature reduction technique: (i) to reduce training time and (ii) to reduce overfitting.

We have used Principal Component Analysis (PCA) because it is one of the most crucial feature reduction techniques. PCA transforms the dataset into a compressed format using linear algebra (Brownlee, 2018). The advantage of using PCA is that users can select the number of dimensions or principal components in the transformed outcome.

- Spatial interpolation

The spatial coverage of GSV images is distributed along with the accessibility of road networks through motorbikes. GSV images help us to visualize the urban scene at each GSV location, i.e., "The GSV images capture the scenes of nearby visual areas instead of single dots in the space" (Cao et al., 2018, p. 6). Therefore it is important to project the extracted ground-level information of GSV images from the bird's eye view using the feature map. The feature maps were generated using the Inverse Distance Weighted (IDW) spatial interpolation technique.

I. FCN Architecture

We have used two different approaches to integrate RSI with the feature maps of SVI, as explained below.

- Approach 1: FCN-DK6-i

The feature maps were stacked with RSI using a composite band tool in ArcGIS result in an increase in the band of the stacked imagery. The stacked imagery was used as an input to FCN-DK6 to classify slums in the study area. As mentioned 5.3.2.1, FCN-DK6 supports n-number of bands. Therefore, FCN-DK6 was used for the stacked imagery. The FCN-DK6 used for stacked imagery is called FCN-DK6-i from now on in the thesis to avoid confusion between FCN-DK6 architecture used for RSI alone and FCN-DK6 architecture used for stacked imagery (RSI + feature maps).

- Approach 2: Modified FCN-DK6

Cao et al. (2018) have used FCN-VGG16 for fusing RSI with SVI for classifying urban land use, as mentioned in Section 2.3.4. The drawback of using FCN-VGG16 is the high computational cost compared to FCN-DK6, and FCN-DK architecture was unexplored for fusing RSI and SVI. Therefore we have proposed Modified FCN-DK6 architecture for fusing RSI with SVI in this research.

We have proposed an innovative approach to integrating the feature map of SVI with RSI using FCN-DK6 architecture with 6 convolutional blocks. However, the feature map of SVI is concatenated with the output of 2nd convolutional block in FCN-DK6 architecture. Figure 5.11 shows the architecture of Modified FCN-DK6 used for this research, and a detailed specification of each block is given in Annexure-I.

We have modified the FCN-DK6 architecture to use two individual inputs to generate a single classification output. The first input was passed through the first 2 convolutional blocks, and then the output of the 2nd convolutional block is concatenated with the second input at the fusion layer. Then the fusion layer is fed to the 3rd convolutional block as an input, and then the output of the 3rd convolutional block is passed through from the remaining convolutional blocks. Finally, the classified map is generated. In this approach, RSI was the first input, and the stacked feature map of SVI was the second input.

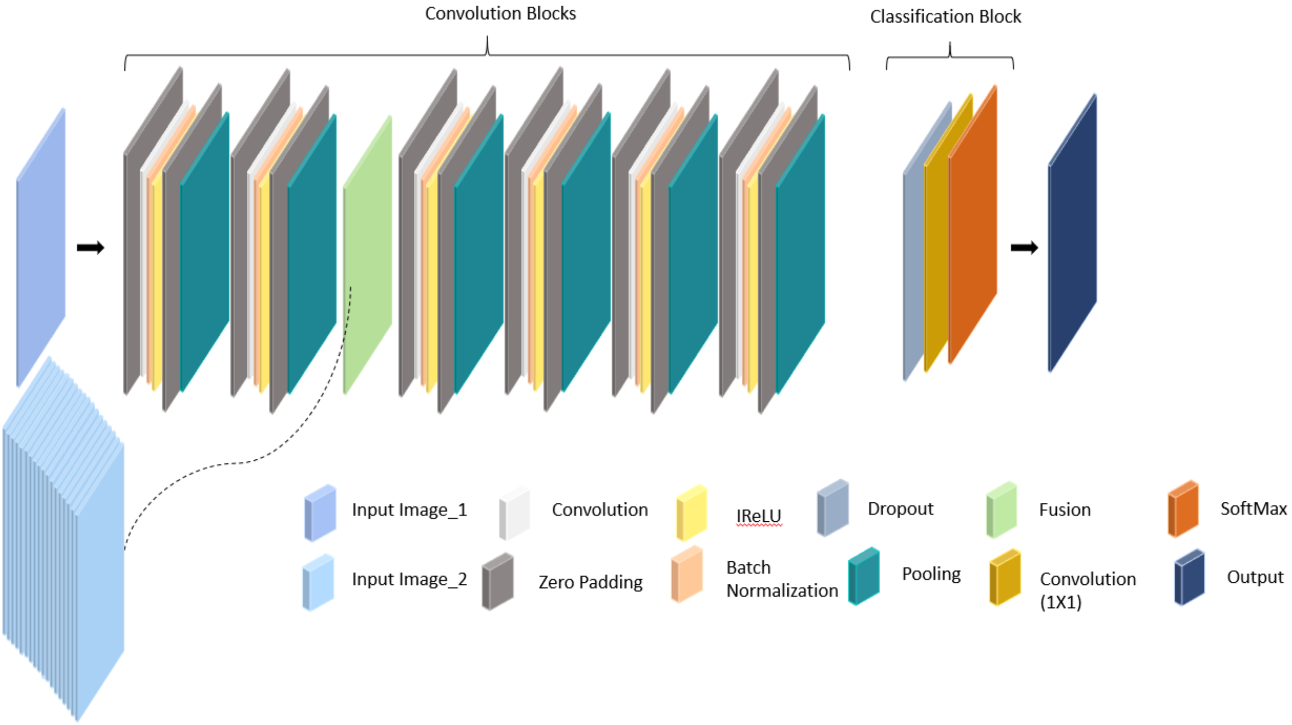


Figure 5.11: Proposed Modified FCN-DK6 architecture for integrating RSI and feature map of SVI (Input 1 is RSI and Input 2 is SVI feature map)

The convolutional block has the trade-off between location and semantic information. The initial blocks have more precise location information, whereas end blocks have more precise semantic information. Therefore we have fused the RSI and stacked feature map of the SVI at the end of 2nd convolutional block to balance location and semantic information.

II. Data Preparation

The data preparation consists of three main steps. The first step discusses the generation of feature maps from SVI. The second step discusses the generation of new points around GSV locations, and the last section discusses the preparation of two different inputs for two different approaches, as discussed above. Figure 5.14 shows the data preparation for Approach 1 and Approach 2.

- Feature extraction

All the selected 39496 GSV images obtained from 9874 GSV locations were used to generate the feature map. The fine-tuned Places365-VGG16 network was used to extract 128 features for each GSV image. Later PCA was used to reduce the dimensionality of extracted features of each image from 128 to 32 because the total variance shown with 32 features is 58.35% of 128 features. Finally, we produced 32 features for each GSV image, i.e., each GSV location has 32 features in each direction (east, west, north, and south) according to their GSV images. Figure 5.12 shows the variance percentage for 32 principal components generated through PCA.

- New point generation

We have generated one point at a distance of 0.5 cm in each direction (east, west, north, and south) for the selected GSV location, i.e., 9874, as shown in Figure 5.13. Then the extracted features are appended to the newly generated points according to their GSV location and the relevant direction. Further feature map has been generated using IDW spatial interpolation technique in ArcGIS. The number of feature map generation depends on how many features are appended to each point. For example, if two features are appended to each point, only two feature maps will be produced. Figure 5.14 shows the procedure for generating the feature maps.

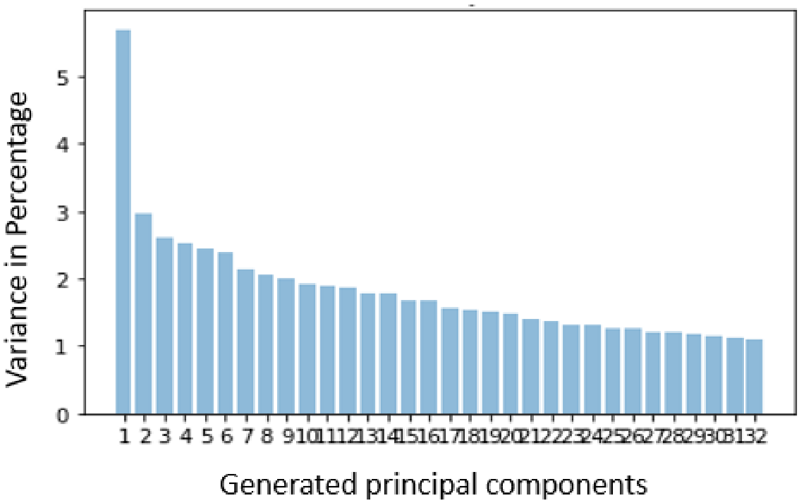


Figure 5.12: Variance percentage for 32 principal components generated through PCA, i.e., 128 features each SVI were reduced to 32 features using PCA

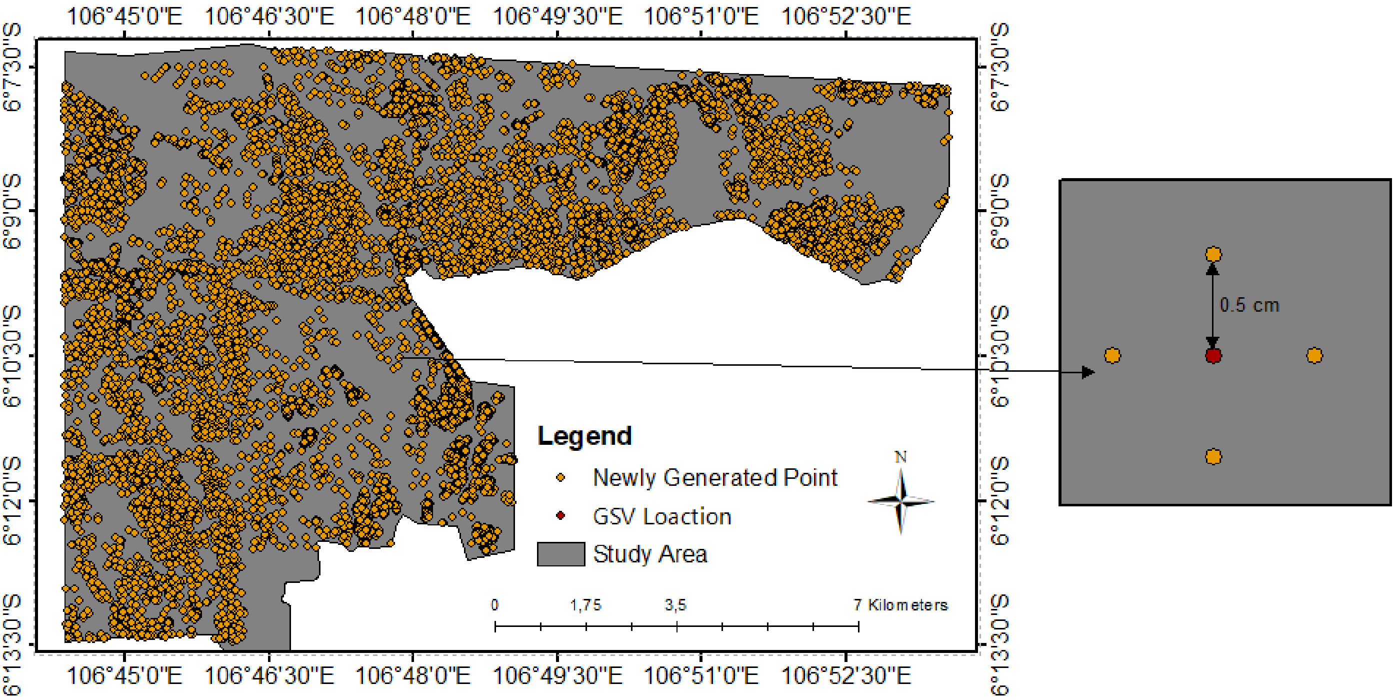


Figure 5.13: New points were generated in the cardinal direction at a distance of 0.5 cm from each GSV location (red are GSV location and yellow are newly generated points)

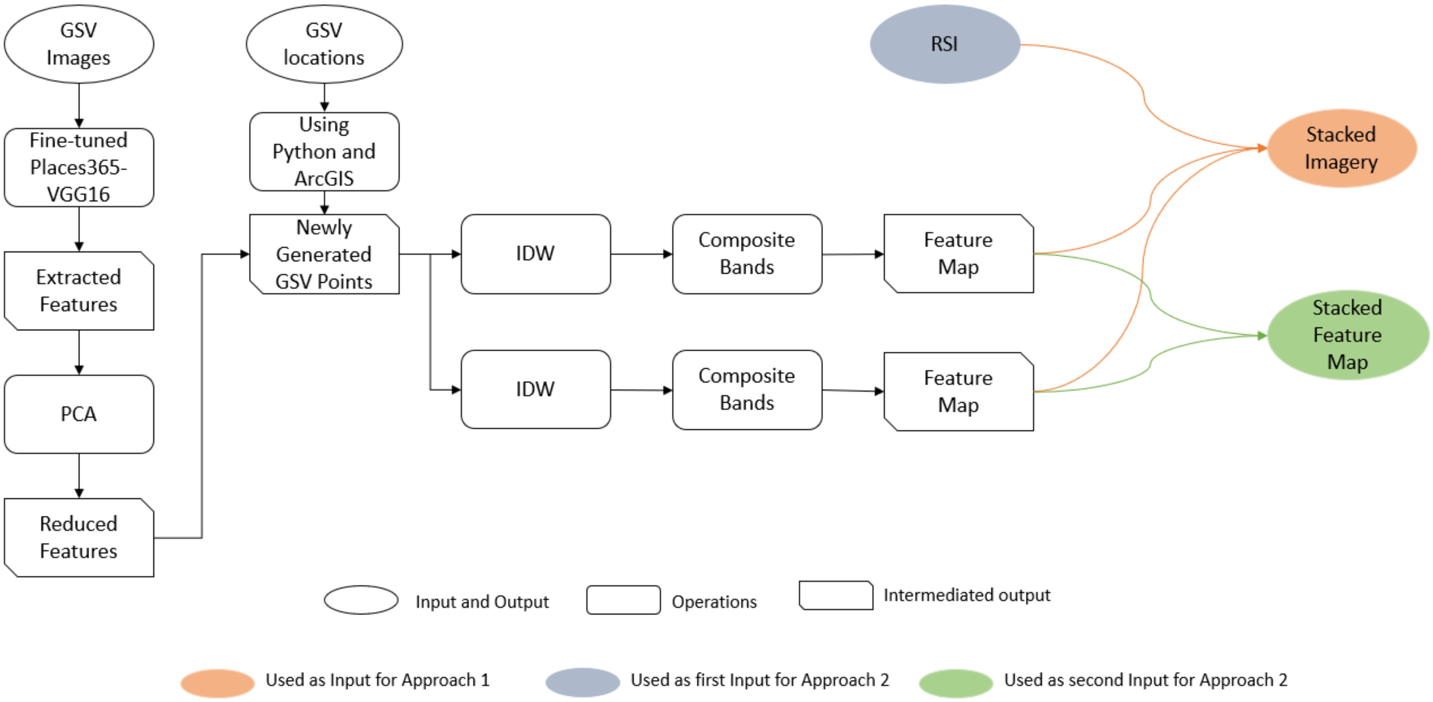


Figure 5.14: Generation of feature maps and the data preparation for Approach 1 and 2
(this figure only shows the data preparation from two features; similarly, it can be done for 32 features)

- Input dataset

Approach 1: The generated feature maps will be stacked with the four bands of RSI using composite bands tools. Then the stacked imagery was clipped according to 12 ground truth tiles generated based on the tweaked slum reference map, as discussed in Section 5.3.1.2. The snapshot of RSI and ground truth tiles prepared for training FCN-DK6-i is shown in Figure 5.15.

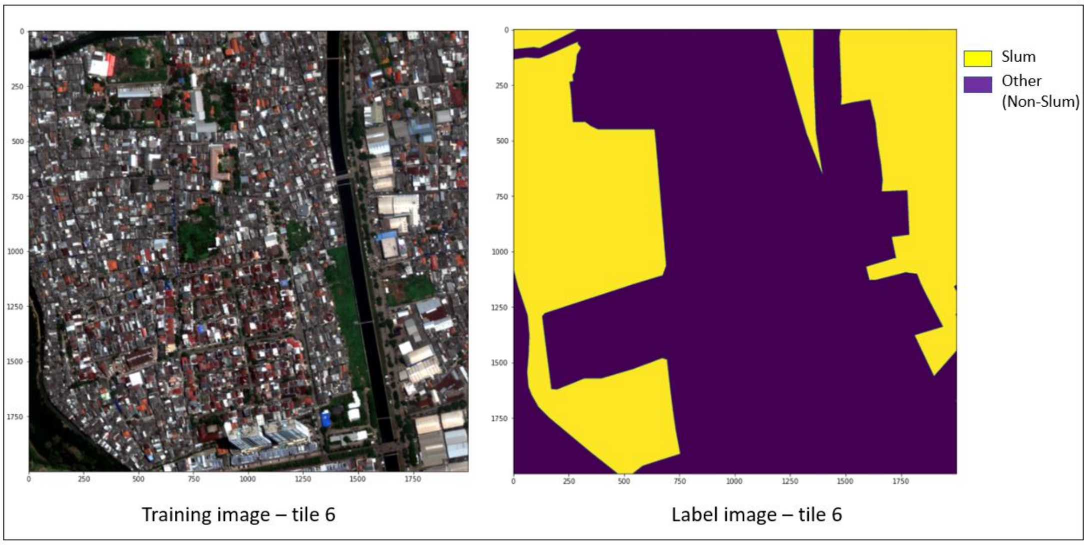


Figure 5.15: Snapshot of RSI and ground truth tiles prepared for training FCN-DK6-i

Approach 2: This approach required two inputs. For the first input, the RSI was clipped according to 12 ground truth tiles generated based on the tweaked slum reference map, as discussed in Section 5.3.1.2. For the second input, the stacked feature map produced by stacking the different feature maps was clipped according to 12 ground truth tiles generated based on the tweaked slum reference map, as discussed in Section 5.3.1.2. The snapshot of the input tiles with the ground truth tiles for training Modified FCN-DK6 is shown in Figure 5.16.

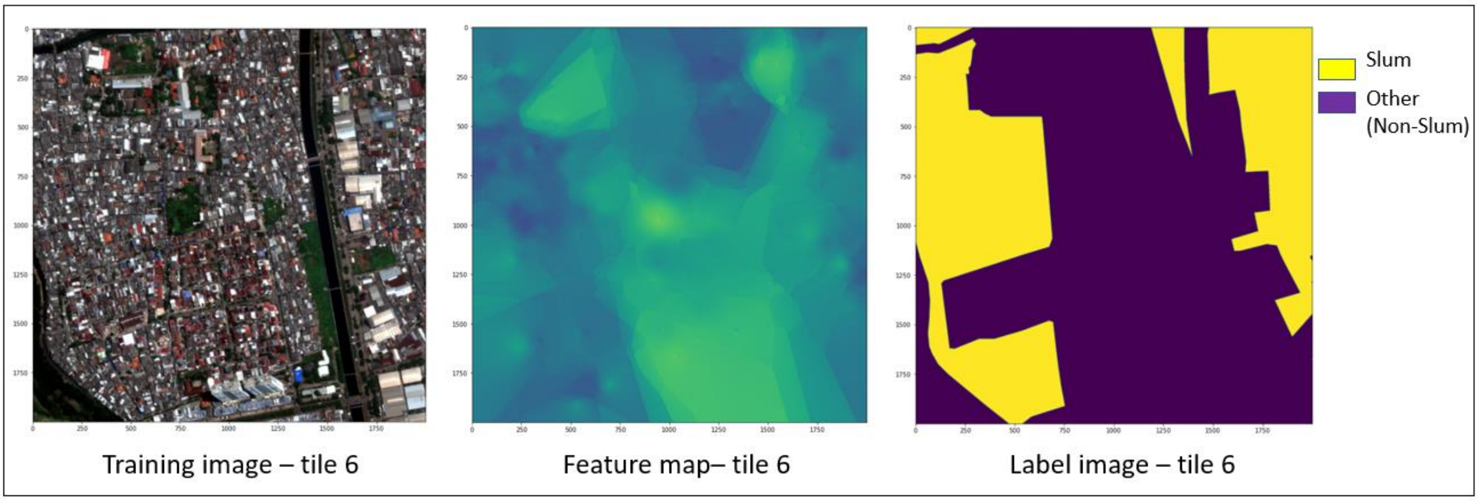


Figure 5.16: Snapshot of the input tiles, i.e., RSI and feature map with the ground truth tile for training Modified FCN-DK6

III. Training Network

- Approach 1: FCN-DK6-i

The samples were systematically extracted from each training tile by splitting them into non-overlapping equal-area patches with the dimension of 125 x 125 pixels. Based on many trials, the size of the patch dimension was fixed. Thus 256 patches have been created for each training tile. Finally, the network configuration is shown in Table 5.6 was used to train FCN-DK6-i, and weights are randomly initiated. We have also used EarlyStopping methods from Keras library to monitor validation accuracy with the patience of 10 epochs, i.e., the model will stop training if it doesn't see any rise in validation accuracy in the past 10 epochs.

Number of epochs	400
Batch size	64
Validation split	0.30
Optimizer	Stocastic Gradiant Desent (SDG) Learning rate: 1x 10^{-5} Momentum: 0.9

Table 5.6: Network configuration used for training FCN-DK6-i in this research

- Approach 2: Modified FCN-DK6

The samples were systematically extracted from each training tile by splitting them into non-overlapping equal-area patches with the dimension of 125 x 125 pixels. Based on many trials, the size of the patch dimension was fixed. Thus 256 patches have been created for each training tile (only for the first input). Finally, the network configuration is shown in Table 5.7 was used to train Modified FCN-DK6, and weights are randomly initiated. We have also used EarlyStopping methods from Keras library to monitor validation accuracy with the patience of 10 epochs, i.e., the model will stop training if it doesn't see any rise in validation accuracy in the past 10 epochs.

Number of epochs	400
Batch size	64
Validation split	0.30
Optimizer	Stocastic Gradiant Desent (SDG) Learning rate: 1x 10^{-5} Momentum: 0.9

Table 5.7: Network configuration used for training Modified FCN-DK6 in this research

5.4. Accuracy Assessment Stage

The prediction accuracy was assessed on the testing tile to evaluate the model's accuracy in this research. As mentioned in Section 2.3.5, different slum mapping researchers have used different accuracy indicators such as kappa coefficient, overall accuracy (OA), recall, precision, F1 score, and IoU. According to Mohammad & Sulaiman (2015), the kappa coefficient is not fit for the image classification problem. Likewise, OA also produces misleading results, especially when classes are imbalanced in the classified image, i.e., a higher OA value doesn't always account for the better performance of the model, such as slum mapping task where non-slum areas dominate slum areas. Therefore this research will use recall, precision, F1 score, and IoU for assessing the model results (Gao, 2020).

Precision is defined as the proportion of correctly classified slum pixels/images to the total classified slum pixels/images, as shown in Equation 1. The recall is defined as the proportion of correctly classified slum pixels/images to the total actual slum pixels/images, as shown in Equation 2. The F1 score is defined as the harmonic mean of precision and recall, as shown in Equation 3. Finally, IoU is defined as the proportion of correctly classified slum pixels/images to the sum of total classified slum pixels/images plus falsely classified slum pixels/images, as shown in Equation 4.

$$P = \frac{TP}{TP + FP} \qquad Eq.1$$

$$R = \frac{TP}{TP + FN} \qquad Eq.2$$

$$F1\ score = \frac{P * R}{P + R} * 2 \qquad Eq.3$$

$$IoU = \frac{TP}{TP + FP + FN} \qquad Eq.4$$

Precision (P), Recall (R), Ture-Positive(TP), False-Positive (FP), True-Negative (TN), False-Negative (FN)

The confusion matrix has been generated for each testing tile where the columns represent the reference data (ground truth), and the rows represent predicted results. First, the total number of ground truth pixels and the predicted outcome pixels were breakdown into binary classes, i.e., slum and other (non-slum), and assigned to the relevant columns and rows as shown in Table 5.8. Then the accuracy of each testing tile is calculated using recall, precision, F1 score, and IoU. Further in this research, the predicted slum pixels were broken down into different categories of slums, and their accuracies were calculated for each slum category.

		Ground Truth	
		Slum	Non-Slum
Predicted Result	Slum	True Positive (TP)	False Positive (FP)
	Non-Slum	False Negative (FN)	True Negative (TN)

Table 5.8: Design of confusion matrix used for summarizing the performance of proposed architectures in this research

5.5. Software and Platform

QGIS was used to generate the random points using random points along line tool, and ArcGIS was used for geospatial operations such as Overlay, Reprojection, Add XY coordinate, Spatial Join, Polygon to Raster was used to generate rasterized slum reference layer as the same spatial resolution of the RSI image, IDW spatial interpolation was used for generating feature maps from the extracted features of SVI, and Composite Band tool was used for stacking RSI and feature maps of SVI with the help of the model builder.

Python was used to implement the different models: FCN-DK6, Places365-VGG16, FCN-DK6-i, and Modified FCN-DK6. The models are based on the Tensorflow 2.x library (with inbuilt Keras library). The models were trained on the Jupyter lab hosted on the "CRIB" (ITC geospatial hub sever). The server has a high computing unit with an inbuilt GPU, i.e., Jetson AGX (8-core ARMv8.2, 32 GB, GPU). The Python script and the relevant data were stored on CRIB, and the satellite imagery was stored on ITC geodatabase as a backup.

6. RESULTS

The results of this research are presented in this chapter. Section 6.1 shows the characteristics of slums in our study area. Section 6.2 shows the predicted outcome of the proposed architectures used in this research. Section 6.3 compares the cumulative accuracy of proposed architectures and compares the accuracy of different slum categories between proposed architectures.

6.1. Identification Outcome

Understanding local slum characteristics in Jakarta is a crucial part of this research. We used governmental documents, academic literature, and ground-level insight of a local expert (Mr. Jati Pratomo) and come up with the list of slum characteristics, which can be identified through RSI, SVI, and ancillary data in our study area to conceptualize the slums in this research, as shown in Table 5.2 and generate tweaked slum map for this research.

As discussed in Section 5.3.1.1, the tweaked reference map consists of two major classes, i.e., slum and other (non-slum). Further, slums were divided into high, medium, light, and very light categories according to the local government of Jakarta, and we defined the unknown slum category. Figure 5.5 shows the tweaked slum reference map used for this research with different slums categories.

6.2. Experimental Outcome

Different architectures were proposed in this research to identify slums using 3 dataset combinations: (1) RSI, (2) SVI, and (3) the combination of RSI and SVI. The results of the proposed architectures are presented below with their corresponding accuracy assessment table. The accuracy of the architecture is calculated based on the ground truth (tweaked reference data) using the accuracy indicators: precision, recall, F1 score, and IoU, as discussed in Section 5.4.

We went one step further to understand our results from the analysis by categorizing the predicted slums into different slum categories based on the tweaked reference map. Thus, we can determine how well different categories of slums were understood from the proposed architectures.

1. RSI

The RSI was used to train the proposed FCN-DK6 architecture with the network configurations shown in Table 5.4. The network is tested on two tiles, and the outcome was a slum map with two classes shown in Figure 6.1 and Figure 6.2, where Figure 6.1 shows the predicted outcome of tile-3 and Figure 6.2 shows the predicted outcome of tile-12. The cumulated accuracy metrics of the testing tiles are presented in Table 6.1, and Table 6.2 shows the cumulated confusion matrix with different categories of slums and other (non-slum).

In Table 6.1, the classification result of FCN-DK6 gets 78.22% precision which means 78.22% of slum pixels are correctly identified from the total predicted pixels of slums. The recall value is 74.42%, i.e., 74.42% of slum pixels are correctly identified from the total number of slum pixels. The F1 score is 76.28%, which combines precision ad recall values. The value of IoU is 61.65% which means 61.65% of the predicted slum map overlaps with the ground-truth data.

Table 6.2 presents the pixel-level breakdown of the predicted outcome of FCN-DK6 into different categories of slums and shows the recall values of different slum categories, i.e., 45.17% of the heavy slum were correctly identified. Similarly, 68.78%, 79.63%, 64.57%, and 76.72% of medium, light, very light, and unknown slums were correctly identified. The recall value of other (non-slum) was 81.39%.

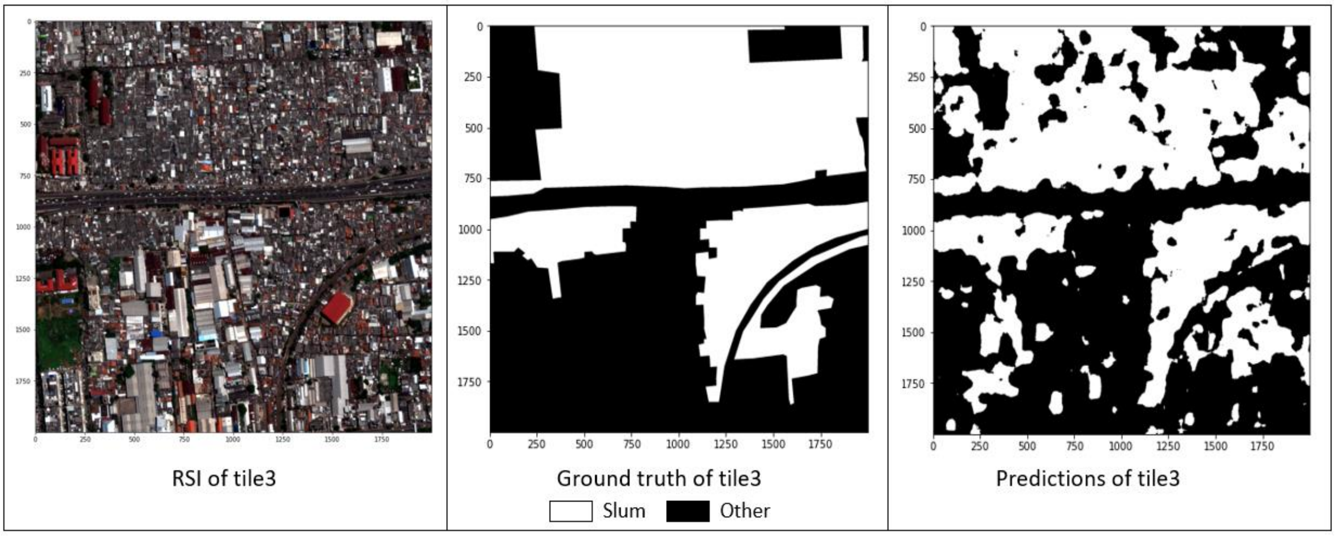


Figure 6.1: Predicted slum map of tile-3 generated from FCN-DK6, where white represents slum and black represents other (non-slum)

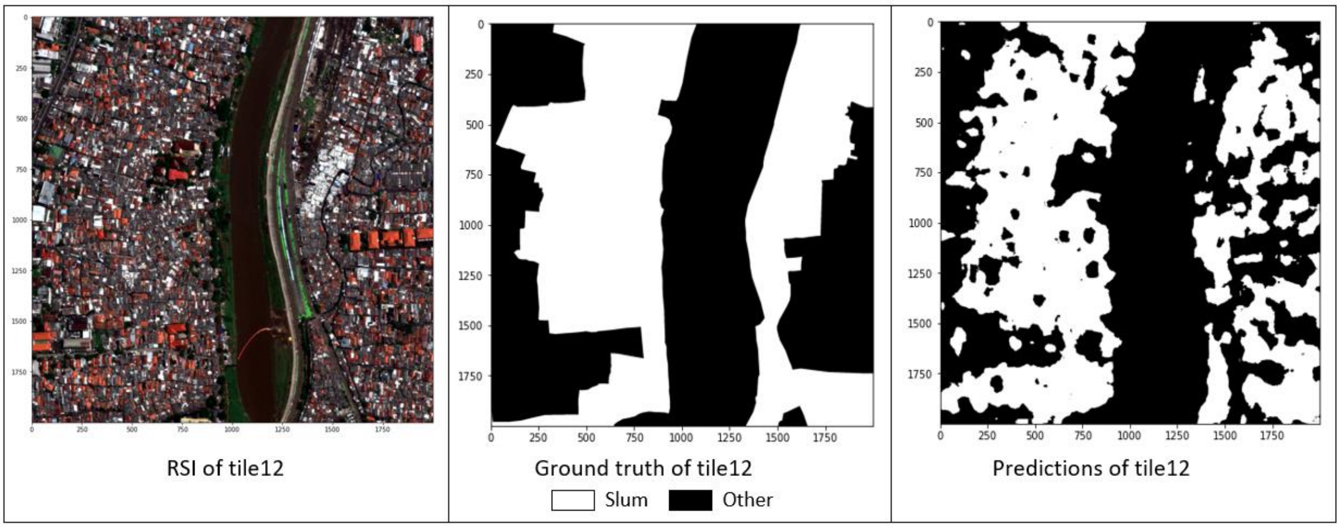


Figure 6.2: Predicted slum map of tile-12 generated from FCN-DK6, where white represents slum and black represents other (non-slum)

FCN-DK6 for RSI	
Precision	78.22
Recall	74.42
F1 score	76.28
IOU	61.65

Table 6.1: Cumulated accuracy metrics of FCN-DK6 are shown in percentage

FCN-DK6 for RSI	Cumulated									
				Ground Truth						Total
				Others (non-slum)	Slum					
					Heavy Slum	Medium Slum	Light Slum	Very Light Slum	Unknown Slum	
	Predicted	Other (non-slum)		3430732	33992	234532	281891	139941	277700	4398788
		Slum	Heavy Slum	784200	28004					3601212
			Medium Slum			516801				
			Light Slum				1102194			
			Very Light Slum					255087		
			Unknown Slum						914926	
	Total			4214932	61996	751333	1384085	395028	1192626	8000000
	Correctly classified in % (Recall)			81.39	45.17	68.78	79.63	64.57	76.72	

Table 6.2: Cumulated confusion matrix of FCN-DK6 are shown with different categories of slums and other (non-slum)

2. SVI

SVI was used to fine-tune the proposed Places365-VGG16 architecture with the network configurations shown in Table 5.5. The network is tested on 7904 images to detect slum and non-slum, i.e., the network is trained to detect the difference between slum and non-slum. The network's output gives the probability of individual images to be categorized as slum or non-slum, shown in Figure 6.3. The accuracy metrics of the network are presented in Table 6.3, and Table 6.4 shows the cumulated confusion matrix with different categories of slums and other (non-slum).

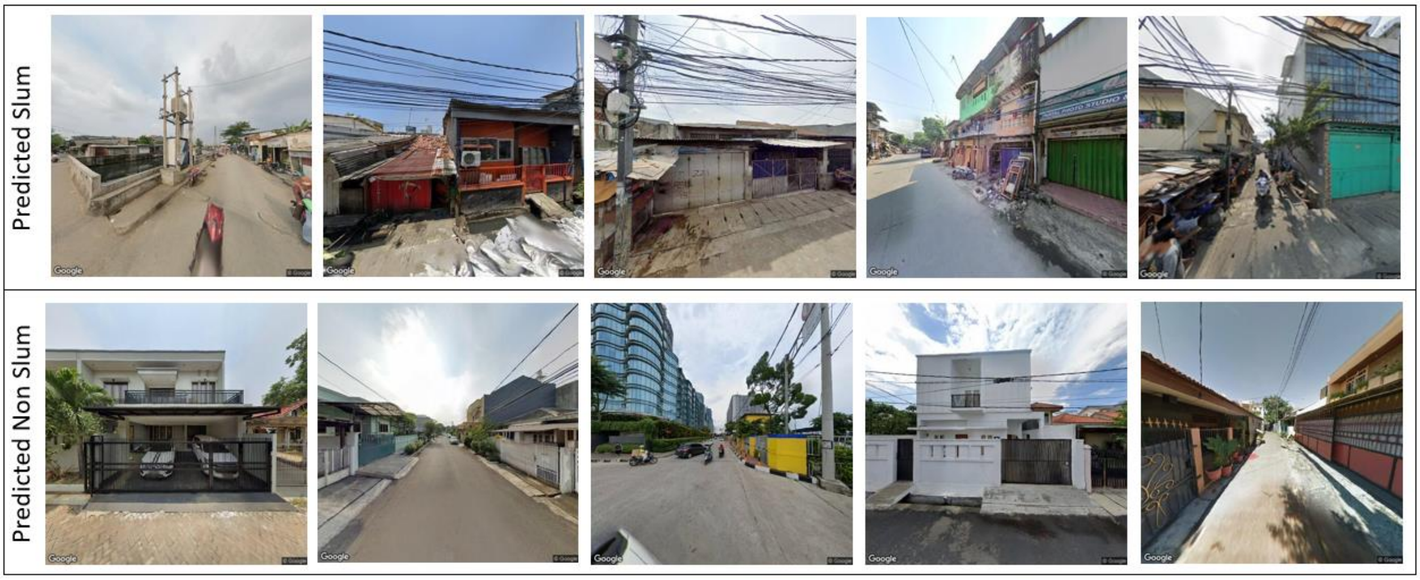


Figure 6.3: Correctly predicted images of slum and non-slum generated from Places365-VGG16

Table 6.3, the classification result of Places365-VGG16 gets 66.66% of precision, .i.e, 66.66% of slum images are identified correctly from total predicted images of slums. The recall value is 76.75%, i.e., 76.75% of slum images are correctly identified from the total number of slum images. The value of the F1 score is 71.35%, i.e., the combination of precision and recall. The value of IoU is 55.46% which means 55.46% of predicted slum images overlap with ground-truth data.

Table 6.4 presents the breakdown of the predicted outcome of Places365-VGG16 into different categories of slums and shows the recall values of different slum categories, i.e., 86.33% of the heavy slum were correctly identified. Similarly, 83.45%, 73.53%, 72.27%, and 74.33% of medium, light, very light, and unknown slums were correctly identified. The recall value of other (non-slum) was 61.61%.

CNN for SVI	
Precision	66.66
Recall	76.75
F1 score	71.35
IOU	55.46

Table 6.3: Cumulated accuracy metrics of Places365-VGG16 are shown in percentage

<table>
<tr><td rowspan="12">CNN for SVI</td><td colspan="10">Cumulated</td></tr>
<tr><td colspan="3" rowspan="3"></td><td colspan="6">Ground Truth</td><td></td></tr>
<tr><td rowspan="2">Other (non-slum)</td><td colspan="5">Slum</td><td>Total</td></tr>
<tr><td>Heavy Slum</td><td>Medium Slum</td><td>Light Slum</td><td>Very Light Slum</td><td>Unknown Slum</td><td></td></tr>
<tr><td rowspan="6">Predicted</td><td colspan="2">Other (non-slum)</td><td>2435</td><td>35</td><td>139</td><td>90</td><td>122</td><td>533</td><td>3354</td></tr>
<tr><td rowspan="5">Slum</td><td>Heavy Slum</td><td rowspan="5">1517</td><td>221</td><td></td><td></td><td></td><td></td><td>4550</td></tr>
<tr><td>Medium Slum</td><td></td><td>701</td><td></td><td></td><td></td><td></td></tr>
<tr><td>Light Slum</td><td></td><td></td><td>250</td><td></td><td></td><td></td></tr>
<tr><td>Very Light Slum</td><td></td><td></td><td></td><td>318</td><td></td><td></td></tr>
<tr><td>Unknown Slum</td><td></td><td></td><td></td><td></td><td>1543</td><td></td></tr>
<tr><td colspan="3">Total</td><td>3952</td><td>256</td><td>840</td><td>340</td><td>440</td><td>2076</td><td>7904</td></tr>
<tr><td colspan="3">Correctly classified in % (Recall)</td><td>61.61</td><td>86.33</td><td>83.45</td><td>73.53</td><td>72.27</td><td>74.33</td><td></td></tr>
</table>

Table 6.4: Cumulated confusion matrix of Places365-VGG16 are shown with different categories of slums and other (non-slum)

3. Two different architectures were implemented for using the combination of RSI and SVI.
   - Approach 1: FCN-DK6-i

The stacked RSI and SVI imagery were used to train the proposed FCN-DK6-i architecture with the network configurations shown in Table 5.6. The network is tested on two tiles, and the outcome was a slum map with two classes shown in Figure 6.4 and Figure 6.5, where Figure 6.4 shows the predicted outcome of tile-3 and Figure 6.5 shows the predicted outcome of tile-12. The cumulated accuracy metrics of the testing tile are presented in Table 6.5, and Table 6.6 shows the cumulated confusion matrix with different categories of slums and other (non-slum).

Table 6.5 was comprehended in the same way as Table 6.1. The classification result of FCN-DK6-i gets 77.38% precision, 75.04% recall, 76.19% F1 score, and 61.54% IoU.

Table 6.6 presents the pixel-level breakdown of the predicted outcome of FCN-DK6-i into different categories of slums and shows the recall values of different slum categories, i.e., 46.79% of the heavy slum were correctly identified. Similarly, 70.23%, 81.43%, 66.25%, and 75.04% of medium, light, very light, and unknown slums were correctly identified. The recall value of other (non-slum) was 80.30%.

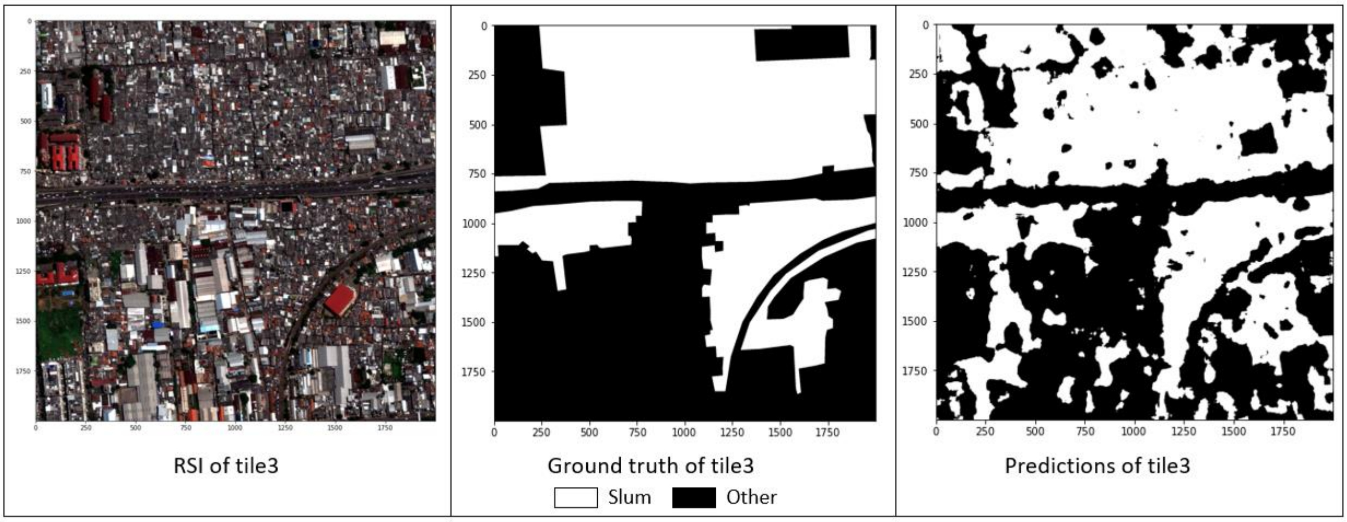


Figure 6.4: Predicted slum map of tile-3 generated from FCN-DK6-i, where white represents slum and black represents other (non-slum)

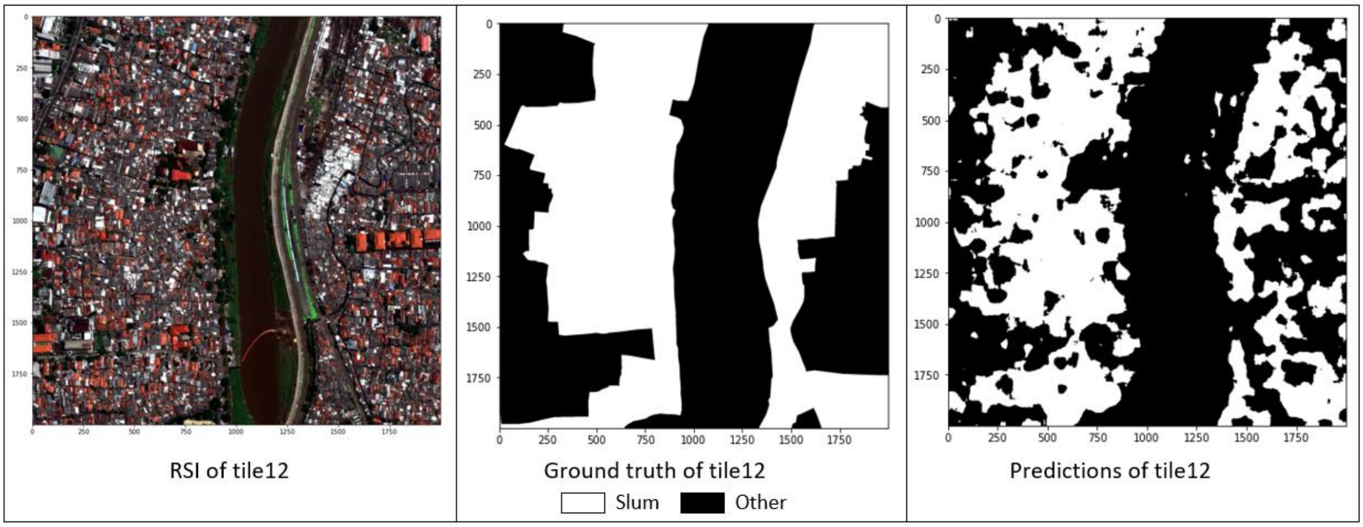


Figure 6.5: Predicted slum map of tile-12 generated from FCN-DK6-i, where white represents slum and black represents other (non-slum)

FCN-DK6-i for Integrated RSI and SVI	
Precision	77.38
Recall	75.04
F1 score	76.19
IOU	61.54

Table 6.5: Cumulated accuracy metrics of FCN-DK6-i are shown in percentage

FCN-DK6-i for Integrated RSI and SVI	Cumulated									
				Ground Truth						Total
				Other (non-slum)	Slum					
					Heavy Slum	Medium Slum	Light Slum	Very Light Slum	Unknown Slum	
	Predicted	Other (non-slum)		3384406	32986	223688	257084	133317	297621	4329102
		Slum	Heavy Slum	830526	29010					3670898
			Medium Slum			527645				
			Light Slum				1127001			
			Very Light Slum					261711		
			Unknown Slum						895005	
	Total			4214932	61996	751333	1384085	395028	1192626	8000000
	Correctly classified in % (Recall)			80.30	46.79	70.23	81.43	66.25	75.04	

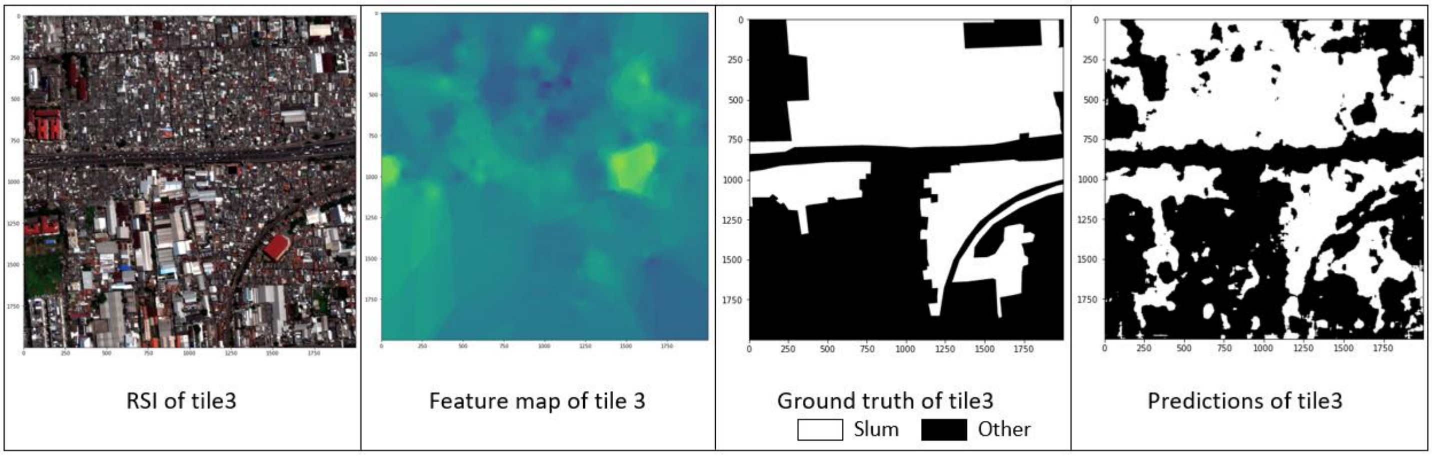
Table 6.6: Cumulated confusion matrix of FCN-DK6-i are shown with different categories of slums and other (non-slum)

- Approach 2: Modified FCN-DK6

Two different inputs, such as RSI and the feature map of SVI, are used to train the proposed Modified FCN-DK6 architecture with the network configurations shown in Table 5.7. The network was tested on two tiles, and the outcome was a slum map with two classes shown in Figure 6.6 and Figure 6.7, where Figure 6.6 shows the predicted outcome of tile-3 and Figure 6.7 shows the predicted outcome of tile-12. The cumulated accuracy metrics of the testing tile are presented in Table 6.7, and Table 6.8 shows the cumulated confusion matrix with different categories of slums and other (non-slum).

Table 6.7 can be comprehended in the same way as Tables 6.1 and 6.5. The classification result of Modified FCN-DK6 gets 77.07% precision, 77.93% recall, 77.50% F1 score, and 63.26% IoU.

Table 6.8 presents the pixel-level breakdown of the predicted outcome of Modified FCN-DK6 into different categories of slums and shows the recall values of different slum categories, i.e., 54.07% of the heavy slum were correctly identified. Similarly, 73.25%, 83.33%, 67.82%, and 79.20% of medium, light, very light, and unknown slums were correctly identified. The recall value of other (non-slum) was 79.18%.



RSI of tile3

Feature map of tile 3

Ground truth of tile3

Predictions of tile3

Figure 6.6: Predicted slum map of tile-3 generated from Modified FCN-DK6, where white represents slum and black represents other (non-slum)

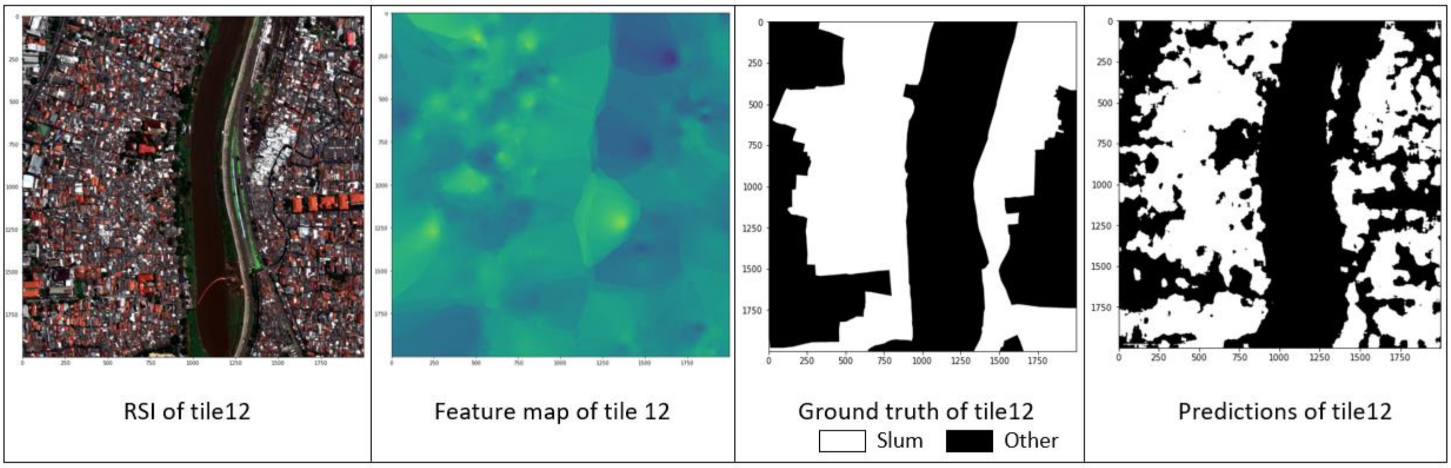


Figure 6.7: Predicted slum map of tile-12 generated from Modified FCN-DK6, where white represents slum and black represents other (non-slum)

Modified FCN-DK6 for Integrated RSI and SVI	
Precision	77.07
Recall	77.93
F1 score	77.50
IOU	63.26

Table 6.7: Cumulated accuracy metrics of Modified FCN-DK6 are shown in percentage

Modified FCN-DK6 for Integrated RSI and SVI	Cumulated									
				Ground Truth						Total
				Other (non-slum)	Slum					
					Heavy Slum	Medium Slum	Light Slum	Very Light Slum	Unknown Slum	
	Predicted	Other (non-slum)		3337202	28475	200990	230783	127138	248024	4172612
		Slum	Heavy Slum	877730	33521					3827388
			Medium Slum			550343				
			Light Slum				1153302			
			Very Light Slum					267890		
			Unknown Slum						944602	
	Total			4214932	61996	751333	1384085	395028	1192626	8000000
	Correctly classified in % (Recall)			79.18	54.07	73.25	83.33	67.82	79.20	

Table 6.8: Cumulated confusion matrix of Modified FCN-DK6 are shown with different categories of slums and other (non-slum)

6.3. Accuracy Outcome

6.3.1. Accuracy Comparison

We look at the F1 score and IoU in more detail instead of looking at precision, recall, F1 score, and IoU for comparing accuracy between the proposed architectures because the F1 score combines precision and recall, and IoU is the benchmark accuracy metric for assessing the PASCAL VOC challenge, which are well-known computer vision competitions (Liu, Qinghui & Salberg, Arnt-Børre and Jenssen, 2018).

The accuracy metrics of predicted outcomes from proposed architectures are represented in Figure 6.8. The horizontal axis shows all the accuracy metrics, i.e., F1 score and IoU. The vertical axis shows the accuracy metrics value in percentage. The different color bars show the accuracy of proposed architectures using different input datasets, i.e., (i) Places365-VGG16 used SVI, (ii) FCN-DK6 used RSI, (iii) FCN-DK6-i used stacked imagery (RSI + SVI feature maps), and (iv) Modified FCN-DK6 used RSI and SVI feature map as two different inputs

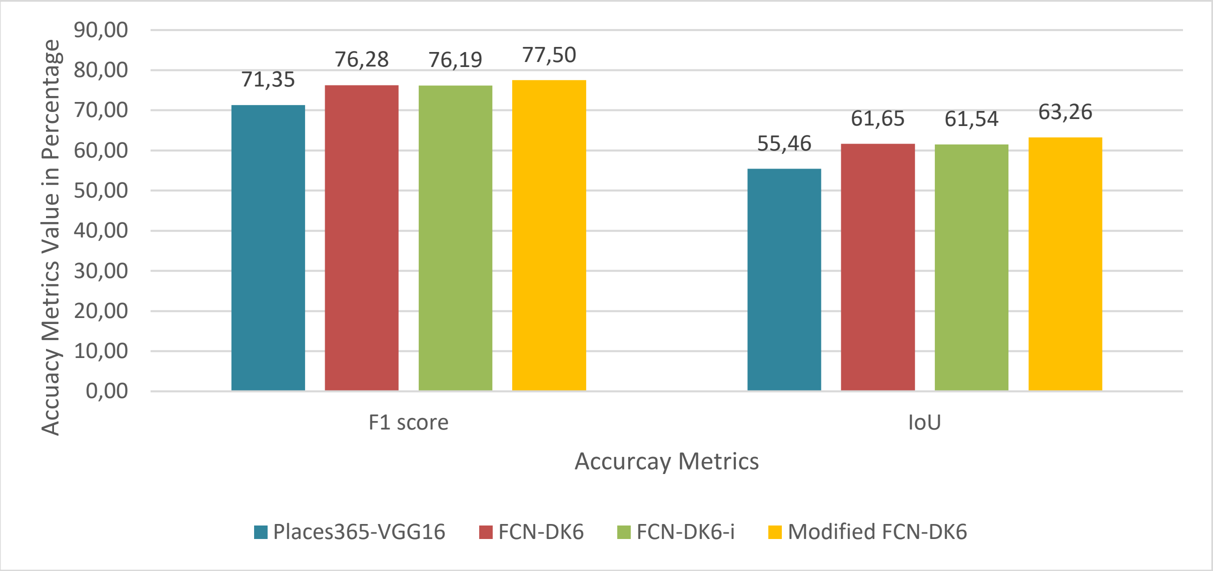


Figure 6.8: Compares accuracy metrics (F1 and IoU) of predicted outcomes from proposed architectures in this research

As shown in Figure 6.8, Modified FCN-DK6 has the highest F1-score and IoU among all the other architectures, which means using the combination of RSI and SVI will lead to better results compared to using RSI or SVI alone. It can be concluded that the help of the ground-level of information from SVI can increase the classification accuracy when combined with RSI. In our case, slums can be mapped more precisely using the combination of RSI and SVI than using RSI or SVI alone.

On the other hand, the FCN-DK6-i also uses the combination of RSI and SVI, but the accuracy is slightly less than the FCN-DK6, which means it is equally important to choose the right network for using the combination of RSI and SVI because the combination of these two datasets won't give a better result with every FCN network.

The proposed FCN-DK6-i is overfilled with input data information, i.e., the stacked imagery has 36 bands such as red, green, blue, near-infrared, and 32 feature maps of SVI. The result generated from FCN-DK6-i using the stacked imagery (high dimensional data) is not so good for two reasons. First, the network couldn't properly understand the features from the input data due to the high dimensionality. Second, edge information plays a critical role in FCN (pixel-wise classification) and there is a high probability of losing the edge information at the time of feature extraction while training the deep learning network with high dimensional data (Liu, Zhenwei, Pan, Zhang, Luo and Lan, 2020).

6.3.2. Slum Categories Comparison

This section compared the breakdown of the predicted outcome into different categories of slums. The categories of predicted slums from proposed architectures are represented in Figure 6.9. The horizontal axis shows different categories of slums. The vertical axis shows the recall values in percentage for

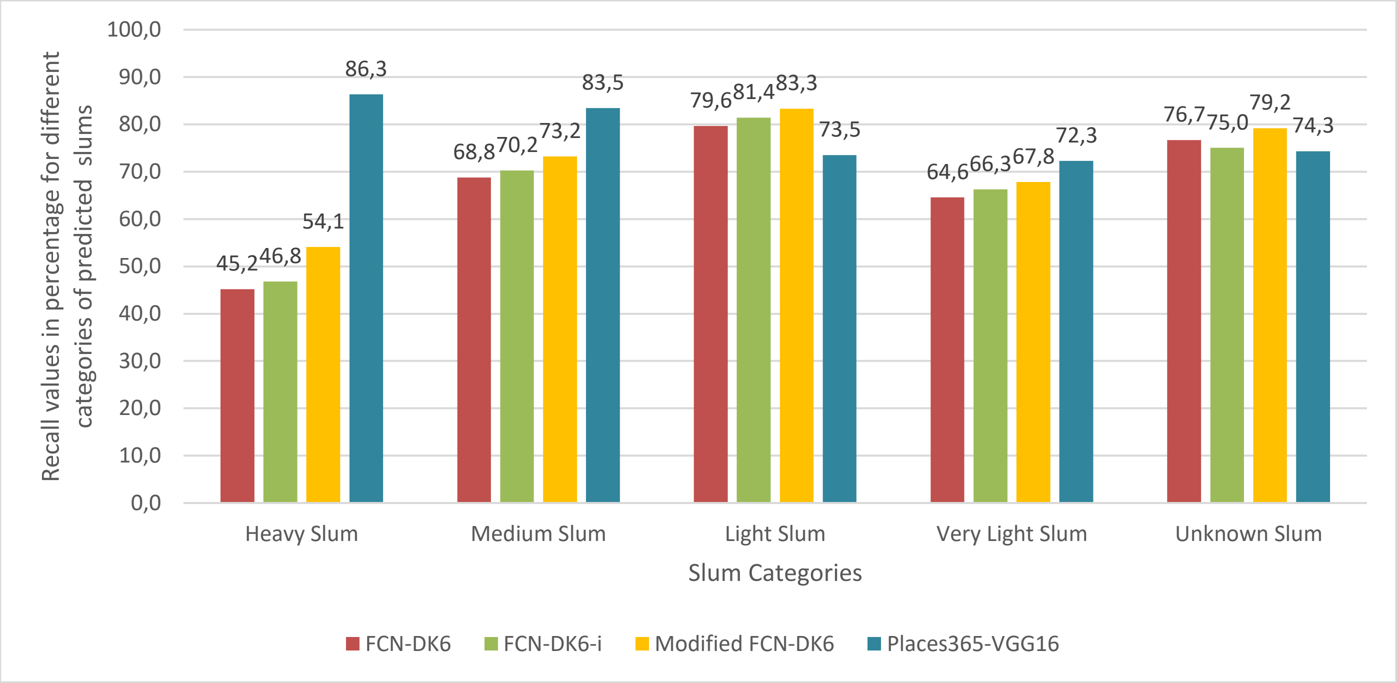


Figure 6.9: Compares predicted outcome into different categories of slums for proposed architectures in this research

different categories of predicted slums. The different color bars show the performance of proposed architectures to identify the different slum categories using different input datasets, i.e., (i) FCN-DK6 used RSI, (ii) FCN-DK6-i used stacked imagery (RSI + SVI feature maps), (iii) Modified FCN-DK6 used RSI and SVI feature map as two different input, and (iv) Places365-VGG16 used SVI.

As shown in Figure 6.9, Places365-VGG16 can identify heavy, medium, and very light slums more precisely than other architectures, but it won't perform well for identifying light and unknown categories of slums. The Modified FCN-DK6 performs better than FCN-DK6-i and FCN-DK6 in identifying the different categories of slums. However, Modified FCN-DK6 performs better than Places365-VGG16 to identify light and unknown categories of slums.

As shown in Figure 6.9, Modified FCN-DK6 and FCN-DK6-i outperform FCN-DK6 for categorizing various slum categories, with the exception of one case where FCN-DK6 outperforms FCN-DK6-i but still won't perform well as compared to Modified FCN-DK6 for classifying an unknown slum category. Hence, the combination of RSI and SVI helps to classify the different categories of slums more precisely than RSI alone.

As shown in Figure 6.9, the SVI understands heavy slums very well than RSI, i.e., the classification percentage of the heavy slum for SVI and RSI is 86.3% and 45.2%. Therefore, there is a significant increase in the classification accuracy of heavy slums in Modified FCN-DK6 compared to FCN-DK6, i.e., from 45.2 % (FCN-DK6) to 54.1% (Modified FCN-DK6).

7. DISCUSSION

This chapter discusses the research outcome. Section 7.1 discusses the characteristics of different categories of slums. Section 7.2 compares the outcome of the proposed architectures based on their performance. Finally, section 7.3 discusses the accuracy of proposed networks.

7.1. Characteristics of Slum

Slum mapping requires a deep understanding of the local characteristics of slums. We have listed different local characteristics of slums in Jakarta using governmental documents, academic articles, and discussion with a local expert (Mr. Jati Pratomo), and discovered the difference between the conceptualization of slums within and between governmental documents and academic articles, making slum mapping challenging. Then we have come up with our definition of slums for this research by selecting a set of slum characteristics, as discussed in Section 5.2.

When we looked into the official slum reference map of 2017 using our definition of slums, we found out that some slum areas were not coinciding with our definition of slums, and some areas that were characterized as slums according to our definition were not delineated as slums in the official slum reference map. Therefore we have tweaked the official slum reference map according to our definition of slums and created a tweaked slum reference map with five categories of slums, as discussed in Section 5.3.1.1. We have used RSI and SVI to understand the different categories of slums. Still, we could not distinguish the different categories of slums using RSI because slum categories possess the same visual characteristics: irregular shape, texture, small building footprint, building density, and proximity to rivers and railroads. In contrast, some ground-level characteristics were found that can be used to differentiate between slum categories by exploring SVI, as shown in Table 7.1. Figure 7.1 shows the ground-level characteristics of slums by exploring SVI.

Ground-Level Characteristics	Description
Inferior building materials	Buildings are constructed with low-quality materials like iron sheets, asbestos sheets, wood, low-grade concrete materials. The low-grade concrete causes cracks, leakages, and improper installation of doors and windows. These problems can be seen with the help of SVI, and unfinished walls of the buildings are also used to identify the inferior building in the study area.
Low-quality roads	The low-quality roads are the roads that are categorized as deteriorated roads and unpaved roads. Although the unpaved roads can be seen through VHR satellite imagery, the low-quality roads can not be seen through VHR satellite imagery.
Open drainage	Uncovered drainage lines along the roads.
Number of floors	An approximate number of floors in the buildings present in that area and the type of construction, i.e., inferior buildings or good quality buildings.
Good building materials	Buildings are constructed with good-quality materials like concrete walls and roofs without cracks, leakages, and improper installation of doors and windows. It can be seen with the help of SVI.

Table 7.1: Explains the ground-level characteristics of slums used for differentiating the slum categories

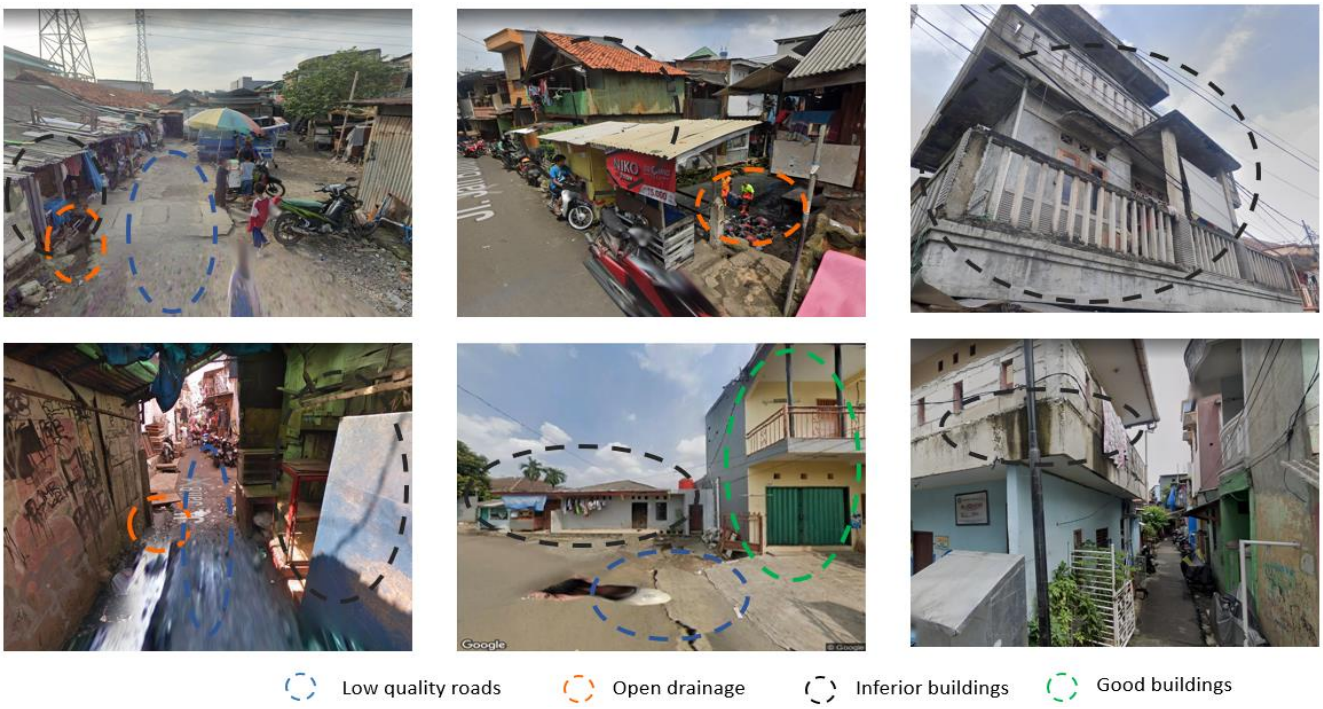


Figure 7.1: Ground-level slum characteristics derived from visual interpretation of SVI in the study area

Ground-Level Characteristics	Heavy Slum	Medium Slum	Light Slum	Very Light Slum	Unknown Slum
Inferior buildings with number of floors	Ground + 1	Ground + 1	Ground + 1	Ground + 1	Ground + 1
Low-quality roads	Yes	Yes	Yes	Yes	Yes
Open drainages	Yes	Yes	Yes	Yes	Yes
Good buildings with number of floors	Mostly Ground floor buildings (few)	Ground floor and Ground + 1 (very few buildings)	Ground floor, Ground + 1 and Ground + 2 (few buildings)	Ground floor, Ground + 1 and Ground + 2	Ground floor, Ground + 1, and Ground + 2 (very few buildings)

Table 7.2: The ground level characteristics observed in SVI corresponding to different categories of slums

By exploring SVI, it can be interpreted that the heavy and medium slum areas can be easily identified compared to other categories of slums (light, very light, and unknown). As shown in Figure 6.9, that fine-tune Places365-VGG16 network understands heavy and medium slum relatively better than RSI. Thus it can be concluded that the heavy and medium slum areas have more general characteristics like inferior building materials, which helps to identify them easily compared to other slums categories, as shown in Table 7.2. In contrast, RSI cannot detect ground-level characteristics like inferior building materials, making it difficult to separate categories based on RSI.

7.2. Applied Architectures

This research used two different types of DL architecture, i.e., CNN and FCN. The CNN architecture was used to categorize each SVI as slum or non-slum. In contrast, FCN architectures were used to generating slum maps. Therefore, this section is divided into two parts. The first part discusses the results generated from CNN architecture, i.e., Places365-VGG16, and the second part discusses the results generated from FCN architectures, i.e., FCN-DK6, FCN-DK6-i, and Modified FCN-DK6.

7.2.1. CNN Architecture

For a better understanding of results generated from the Places365-VGG16 network, we have visualized the features which were used by the network to reach its decision, i.e., slum or other (non-slum). Due to the time constraints, we have only visualized features responsible for classifying the image into slum, i.e., only correctly classified slum images by Places365-VGG16 network were used for visualizing the features, as shown in Figure 7.2.

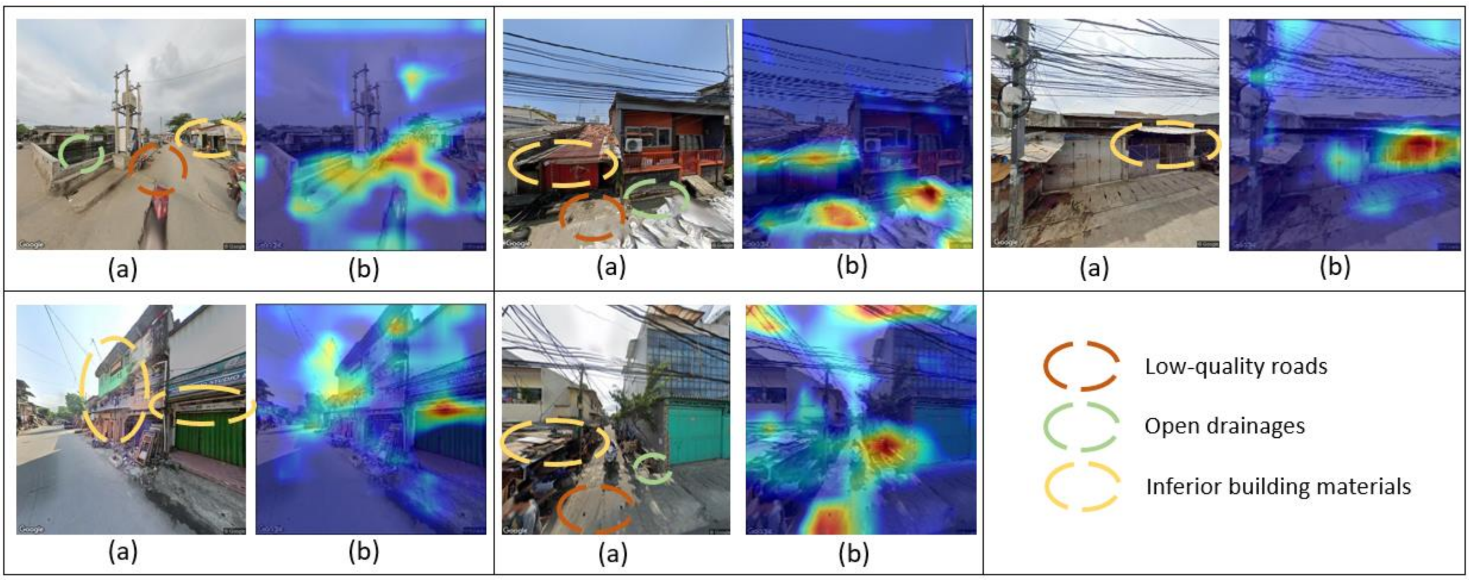


Figure 7.2: The result of Places365-VGG16 for identifying slums, (a) the actual SVI of correctly classified slum images by Places365-VGG16 network, (b) visualize feature map of correctly classified slum images.

Figure 7.2 shows the result of Places365-VGG16 for identifying slums, (a) the actual SVI of correctly classified slum images, (b) visualize feature map of correctly classified slum images. It is evident from Figure 7.2 that the Places365-VGG16 network use ground-level characteristics such as low-quality roads, open drainages, and inferior building materials to identifying slums in an urban scene. Although we did not visualize the features for all correctly classified slum images, we have visualized less than 10% of correctly classified slum images. Therefore, we are not concluding that the ground-level features mentioned above are the only ones responsible for classifying the images into slums. There might be a possibility of finding other ground-level features when more correctly classified slum images will be visualized.

7.2.2. FCN Architectures

We have compared the output generated from different FCN networks in this section because the main objective of this research is to use two different datasets, i.e., RSI and SVI, to map slums in the urban scene.

The FCN-DK6 used RSI alone, whereas FCN-DK6-i and Modified FCN-DK6 used the combination of RSI and SVI to map slums. Adding SVI with RSI increased the understanding of the urban scene better as compared to RSI alone. Figure 7.3 shows the classification of proposed architectures for identifying slum and non-slum, (a) the actual image of slum and the non-slum area with GSV locations, but there is one non-slum building in slum area highlighted with red, (b) FCN-DK6 shows the false prediction of the non-slum building as a slum, (c) FCN-DK6-i shows the better result than FCN-DK6, (d) Modified FCN-DK6 shows the correct prediction of the non-slum building as non-slum and additionally it also increases the coverage of slum area accurately. The Modified FCN-DK6 performs well because of the availability of SVI at that location which helps to improve the prediction. It can be concluded that the ground-level information provided by SVI help to understand the complexity of the urban areas, like how the Modified FCN-DK6 identifies the single non-slum building in a slum area, which was quite hard to identify through FCN-DK6 because of the absence of the ground-level information in RSI. Thus by looking into feature map of the SVI, it can be said that the inferior building material help to differentiate non-slum building in slum area.

In contrast, Figure 7.4 shows the classification of proposed architectures for identifying the non-slum area, (a) the actual image of slum and non-slum area with GSV locations, and (b), (c), and (d) FCN-DK6, FCN-DK6-i, and Modified FCN-DK6 shows the false prediction of the non-slum area as a slum. The proposed architectures show similar results because the non-slum area has similar characteristics as slum areas like irregular shape, small building footprint, high-density buildings, and similar roofing materials as like slum. Those non-slum areas do not have the SVI coverage, due to which FCN-DK6-i and Modified FCN-DK6 show false predictions.

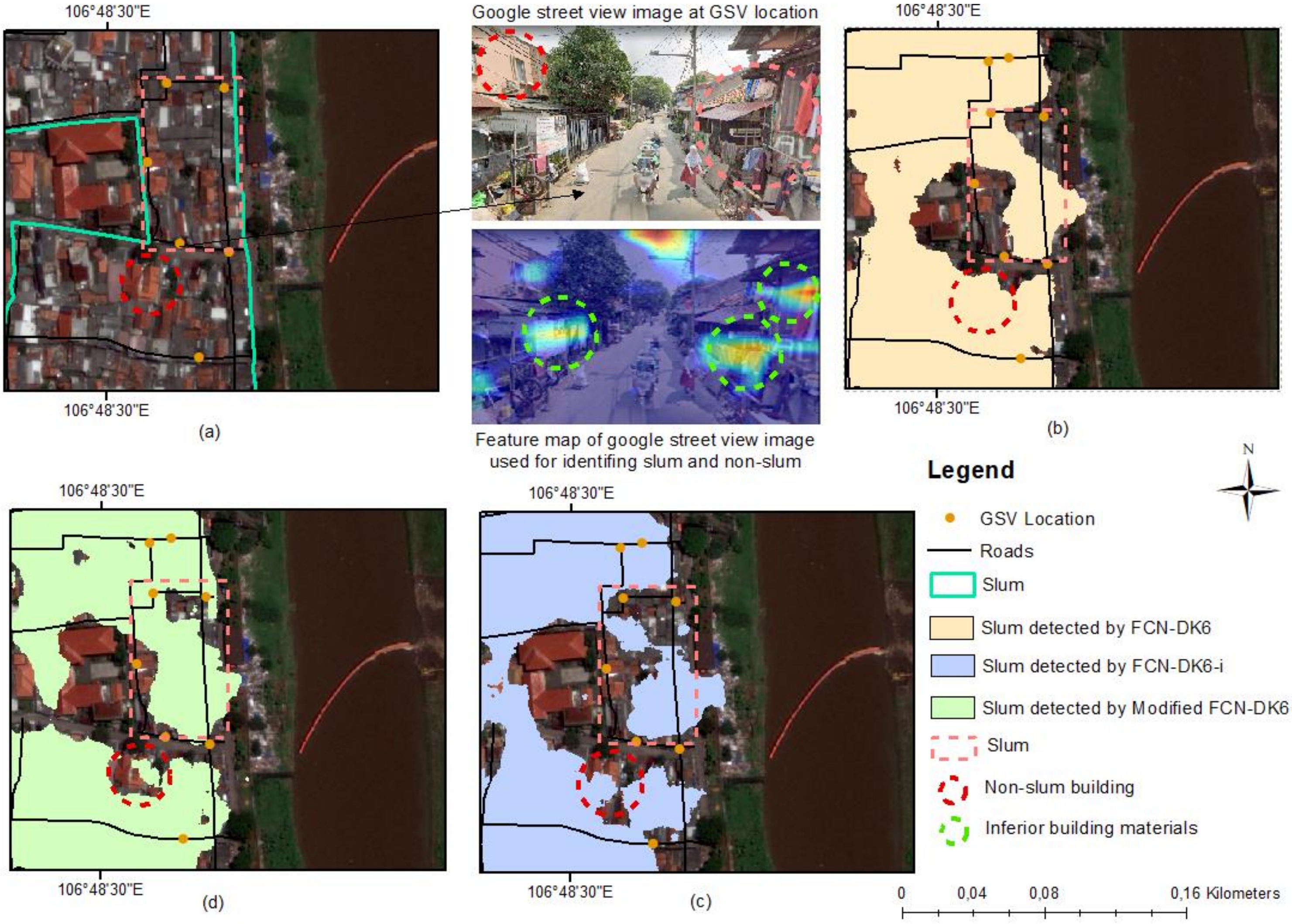

106°48'30"E
Google street view image at GSV location
Feature map of google street view image used for identifing slum and non-slum
(a)
(b)
(c)
(d)
Legend
N
GSV Location
Roads
Slum
Slum detected by FCN-DK6
Slum detected by FCN-DK6-i
Slum detected by Modified FCN-DK6
Slum
Non-slum building
Inferior building materials
0
0,04
0,08
0,16 Kilometers

Figure 7.3: Classification of proposed architectures for identifying slum and non-slum area, the Modified FCN-DK6 identify the non-slum building in a slum area with the help of GSV imagery feature, i.e., inferior building materials

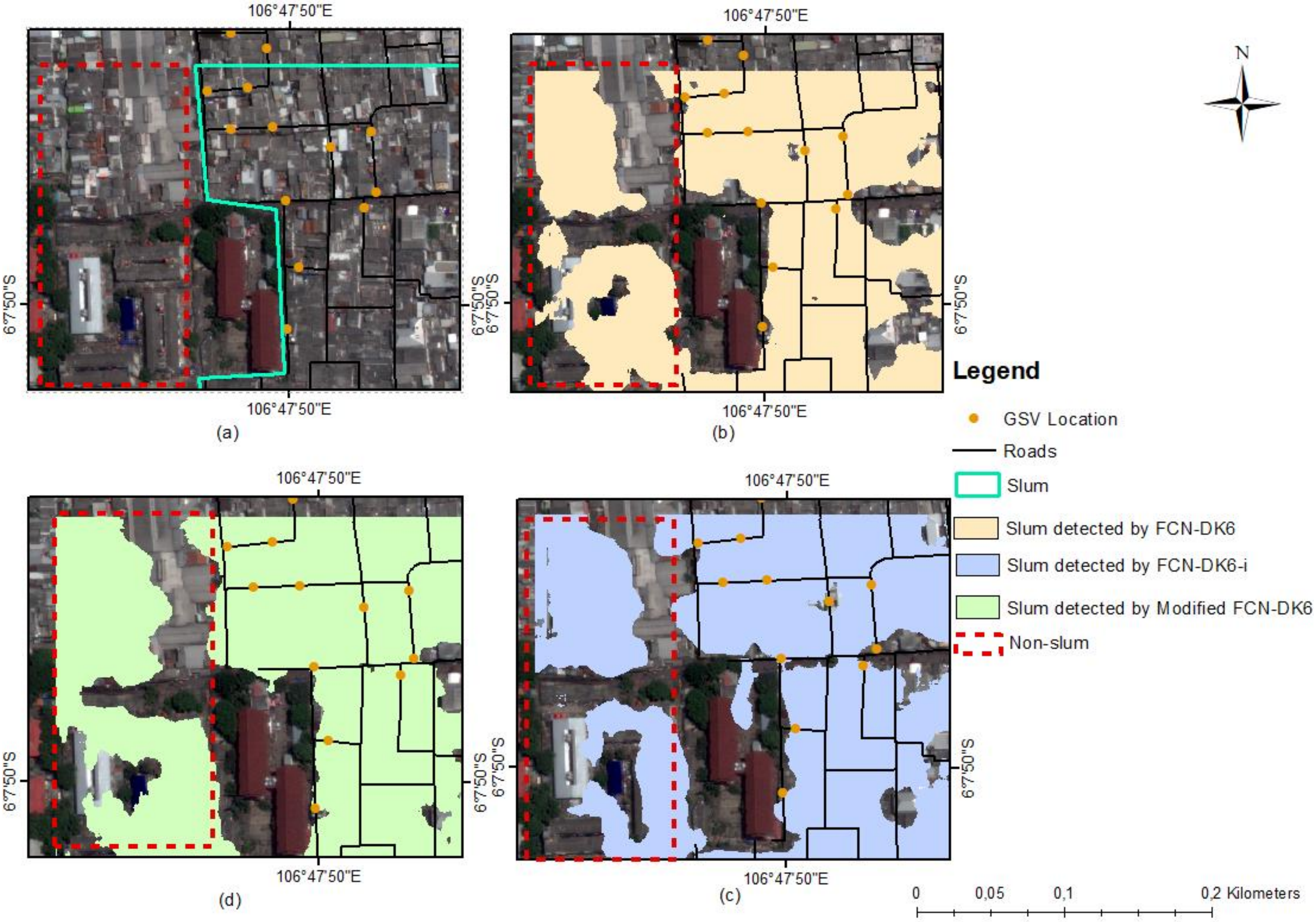


Figure 7.4: Classification of proposed architectures for identifying the non-slum area, all the architectures performed same due to unavailability of GSV location in non-slum area

Figure 7.5 shows the classification of proposed architectures for identifying slum area, (a) the actual image of slum area with limited GSV locations, and (b), (c), and (d) FCN-DK6, FCN-DK6-i, and Modified FCN-DK6 shows the false prediction of slum area as non-slum. The proposed architectures show similar results because slum area didn't show characteristics like slums, especially the roofing material. The FCN-DK6 shows the whole area as non-slum, whereas the FCN-DK6-i and Modified FCN-DK6 predict very few buildings as a slum in that area. The poor performance of FCN-DK6-i and Modified FCN-DK6 is because there is only one SVI available in that area. In contrast, Figure 7.6 shows the classification of proposed architectures for identifying slum area, (a) the actual image of slum area with the GSV locations, (b) FCN-DK6 shows the poor prediction of slum area, (c) FCN-DK6-i shows the better result when compared to FCN-DK6 and (d) Modified FCN-DK6 show the good prediction of slum area. The availability of SVI increases the performance of Modified FCN-DK6 when compared to FCN-DK6. The FCN-DK6-i didn't perform well because of the lack of understanding of the urban scene compared to Modified FCN-DK6.

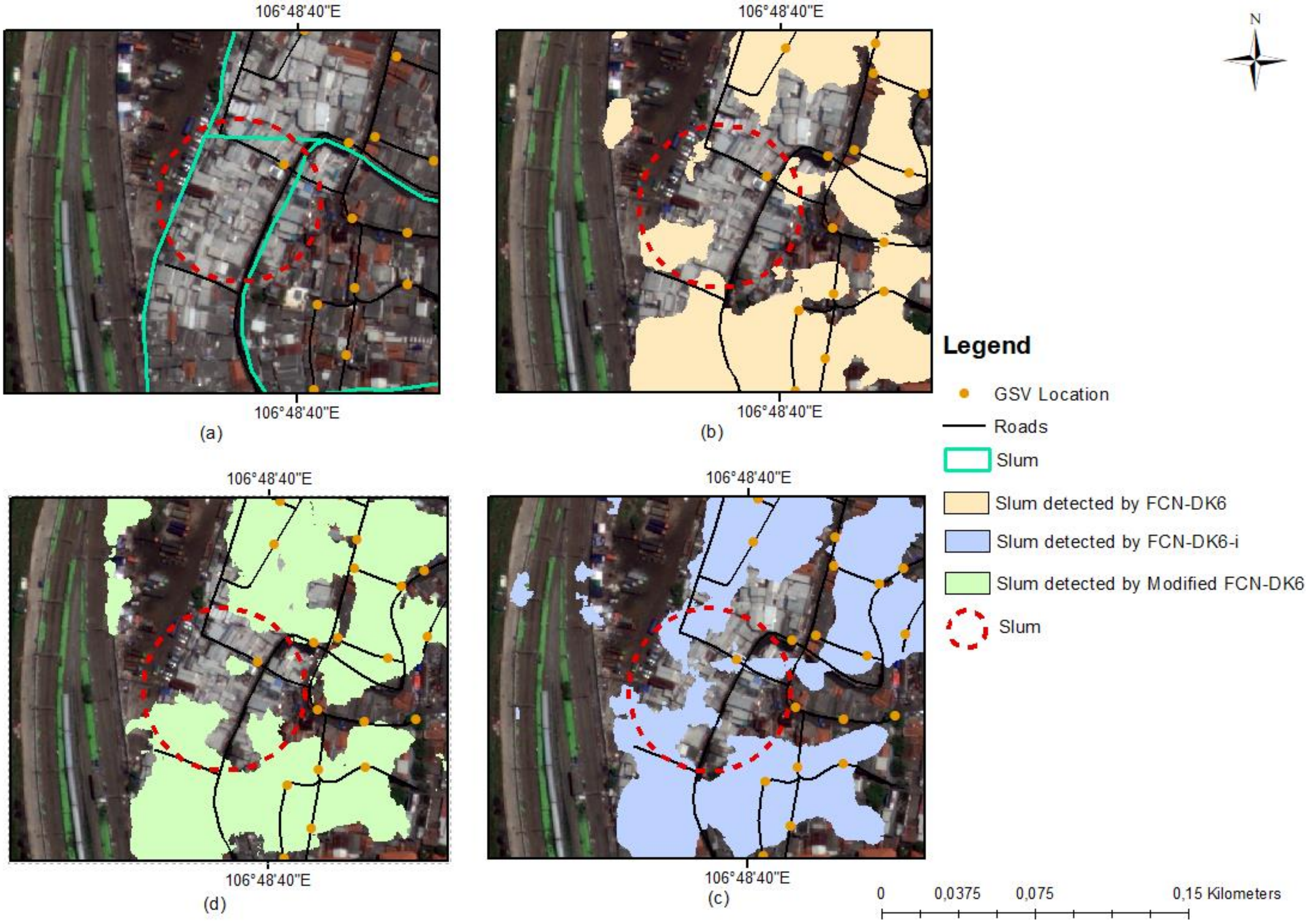


Figure 7.5: Classification of proposed architectures for identifying slum area, where FCN-DK6-i and Modified FCN-DK6 perform poor due to very limited access to GSV location

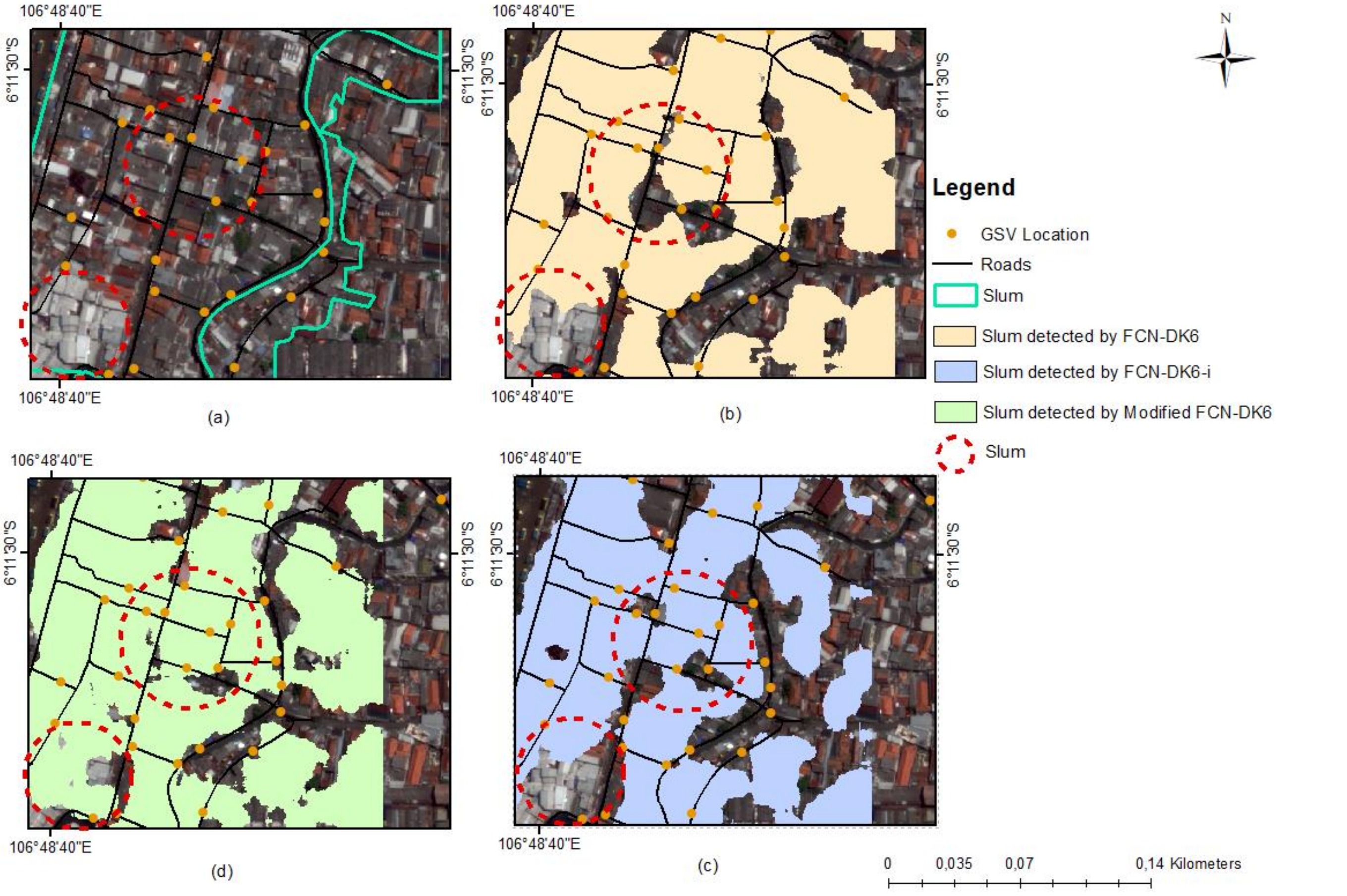


Figure 7.6: Classification of proposed architectures for identifying slum areas where Modified FCN-DK6 perform quite good as compares to others due to the availability of GSV locations

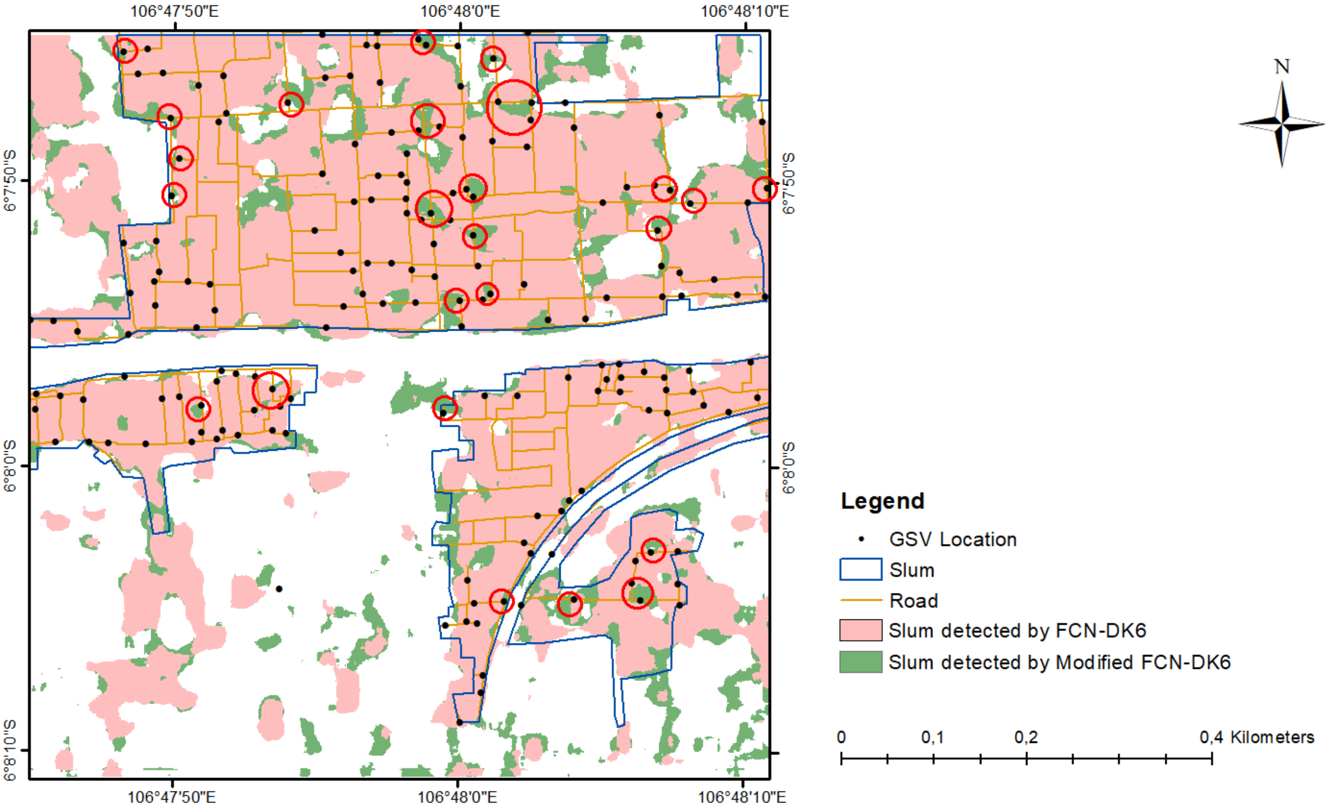


Figure 7.7: Compare predicted slum map of Modified FCN-DK6 and FCN-DK6 for tile-3
The red circles show areas where the Modified FCN-DK6 performs better than FCN-DK6 due to SVI's availability

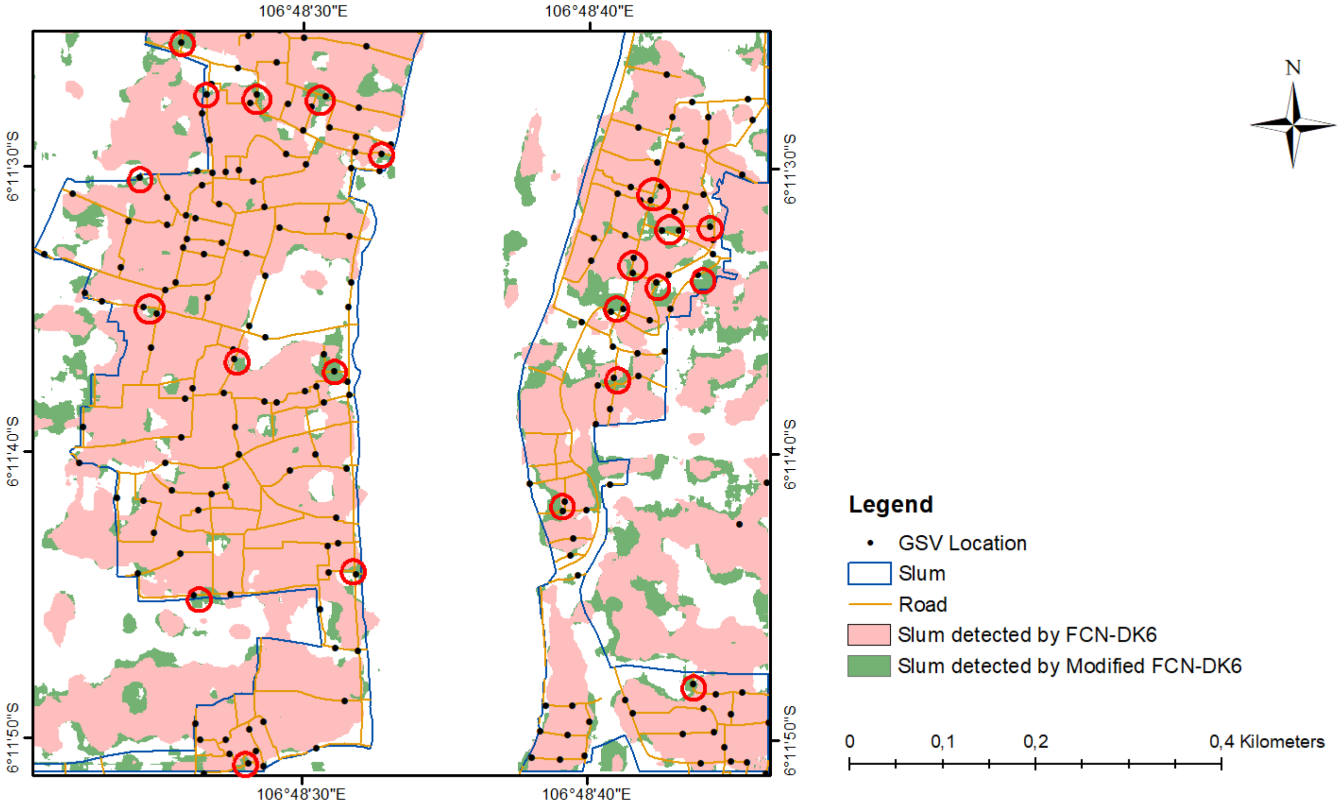


Figure 7.8: Compare predicted slum map of Modified FCN-DK6 and FCN-DK6 for tile-12
The red circles show the areas where the Modified FCN-DK6 performs better than FCN-DK6 due to SVI's availability

Figure 7.7 and Figure 7.8 compares Modified FCN-DK6 and FCN-DK6 outcomes for tile-3 and tile-12. Modified FCN-DK6 shows a better prediction of slums as compared to FCN-DK6. The prediction accuracy of Modified FCN-DK6 increases with the availability of SVI because the SVI provides ground-level information, which helps to understand slums in an urban environment. As shown in Figure 7.7 and Figure 7.8, the red circles show areas where the Modified FCN-DK6 performs better than FCN-DK6 due to SVI's availability.

7.3. Accuracy of Applied Architectures

7.3.1. Accuracy

The ground feature map generated from the SVI provides the ground-level information, which helps improve the accuracy in the prediction when the combination of SVI and RSI is used. However, there is a slight increase in the resulted accuracy of Modified FCN-DK6 as compared to FCN-DK6. There are various reasons why the results didn't change drastically despite a slight increase in the result accuracy, including:

i. The limited coverage of SVI because of two reasons. First, the SVI can only capture the information along the road, making the information limited, i.e., SVI can capture the limited scene along the sides of the roads. Second, some of the roads for which the SVI is not available means the ground-level information is not accessible, i.e., the GSV images are not taken along those roads.

ii. Some information loss might happen while generating feature maps from the extracted features of SVI using the spatial interpolation technique.

iii. It might be possible that the selected 32 features for generating feature maps didn't have sufficient ground-level information to improve the result drastically when it is integrated with RSI for slum mapping. The 32 features show only 58.35% variance of 128 extracted features. Whereas at the time of experimentation, we initially used only 2 features to generate the feature maps instead of 32 features. The result generated using 2 features was quite similar to RSI because 2 features show an 8.2% variance of 128 extracted features. After increasing the number of features, the classification accuracy was increased compared to RSI. Thus, increasing the features will increase the ground-level information, which will finally increase the accuracy of the classification result.

7.3.2. Slum Categories

A combination of SVI with RSI helps identify the different categories of slums more precisely than RSI alone. Figure 6.9 shows that the integration of RSI and SVI leads to an increase in the classification result for every category of slums compared to RSI alone, i.e., Modified FCN-DK6 and FCN-DK6-i understand the different categories of slums better when compared to FCN-DK6. There are two main reasons why there is an increase in the classification value for each category of slums:

i. Some of the detailed ground-level information possessed by SVI helped to differentiate between different categories of slums, as mentioned in Table 7.1 and discussed in Section 7.2.1, because it is hard to get such details from RSI alone.

ii. As shown in Figure 6.9, the fine-tuned Places365-VGG16 network understands the features of different slums categories, especially heavy and medium slum. The Places365-VGG16 network is used to extract SVI features, further combined with RSI for slum classification. Therefore, if the network understands the features of slum properly, then the extracted features from the network will closely represent slum features, which will further help to improve the classification result when combined with RSI for slum mapping. In our case, slum images on which Places365-VGG16 was previously trained consist of similar characteristics as of heavy and medium slum, i.e., inferior building materials, low-quality roads, high-density housing. Therefore Place365-VGG16 identifies the heavy and medium slums quite well. Similarly, integrating RSI with extract SVI features gives better accuracy for heavy and medium slums than other categories of slums.

8. CONCLUSION AND RECOMMENDATIONS

This research aims to integrate RSI and SVI to map slums using a DL approach. Section 8.1 provides the research conclusion by answering the main objective through three sub-objectives with their corresponding research questions. Section 8.2 presents the limitations of this research and recommendations for future research work.

8.1. Conclusion

Slum maps play a significant role in promoting positive changes in national or local upliftment policies. It also helps track down the progress of the implemented upliftment policy and also supports the local people in negotiations with governments about basic services, and asserts their rights to land on which they live. Traditional slum mapping can be labor-intensive and expensive work. We have tried to explore a novel way by integrating RSI with SVI for slum mapping by implementing a DL approach. The results show that the integration of RSI with SVI can achieve relatively higher accuracy than RSI alone, which means that the ground-level information extracted from SVI plays a significant role in identifying slums in an urban area. This research opens a new path for exploring future ways to combine RSI with SVI for slum mapping to achieve better accuracy.

In this research, we trained and tested four networks with different datasets, i.e., FCN-DK6 used RSI alone, Places365-VGG16 was fine-tuned using SVI, and FCN-DK6-i and Modified FCN-DK6 used a combination of RSI and SVI on the subset of Jakarta. We discovered two things: (1) Places365-VGG16 performs well for identifying heavy and medium categories of slums than RSI alone because heavy and medium slum areas have more general ground level characteristics like inferior building materials, low-quality roads, and open drainages, which makes them easy to identify as compared to other categories of slums, and (2) Modified FCN-DK6 outperforms other FCN networks in slum mapping and shows that the combination of RSI and SVI can perform better than RSI alone because SVI consists of useful ground-level information used to classify slums in the urban scene. In contrast, FCN-DK6-i also used the combination of RSI and SVI but achieved slightly less accuracy than FCN-DK6. Thus, it can be concluded that the way how SVI is combined with RSI is a crucial step because FCN-DK6-i and Modified FCN-DK6 used the combination of RSI and SVI. Still, the way of combining SVI with RSI is different in both the networks, i.e., the FCN-DK-i used stacked imagery (RSI + SVI) as one input. In contrast, Modified FCN-DK6 used two different inputs, i.e., the first input was RSI, and the second input was SVI features. The SVI features were combined with RSI at the end of 2nd convolutional block in Modified FCN-DK6.

Research sub-objective and related research questions:

I. To identify the characteristics of slums versus non-slum in the study area.

In this research, the characteristics of slums were shortlisted by reviewing the different governmental organization documents at global, national (Indonesia), and local (Jakarta) scales, research articles at the local level (Jakarta), and discussion with the local expert (Mr. Jati Pratomo). Finally, characteristics were selected from the shortlisted characteristics, which can be delineated using RSI or SVI or available ancillary data such as road network data, building footprint data, and zoning data in our study area.

1. What are the physical characteristics of slums in the study area?
2. Which features should be extracted from RSI to classify slums?
3. Which visual features should be extracted from SVI to classify slums?

Table 8.1 shows the physical characteristics of slums in the study area and which physical characteristics can be captured through RSI or SVI or both.

Physical characteristics of slums in the study area	RSI	SVI
Temporary building materials	No	Yes
Dense area with lesser roads	Yes	No
Unplanned layout	Yes	No
Unpaved/Light roads	Yes	Yes
Small building size/building footprint	Yes	No
Poor roof materials	Yes	Yes (Partially)
Proximity to river, railroads, swamps, and shrines	Yes	Yes
Near to industrial and warehouse area	Yes	No
Less open and green spaces	Yes	Yes (Partially)
Open drainages	No	Yes
Low-quality roads	No	Yes

Table 8.1: The physical characteristics of slums in the study area, which can be captured through RSI or SVI or both

II. To incorporate SVI with RSI for slum mapping using FCN

In this research, two different deep learning approaches were used for identifying slums, i.e., FCN and CNN, i.e., FCN is a pixel-level classification technique, and CNN is a patch-based classification technique. Therefore to do pixel-level classification, FCN was used to incorporate SVI with RSI for slum mapping.

1. Which FCN architecture is the best fit for using the combination of RSI and SVI to identify slums?

The FCN-DK architecture is used because of three main reasons. First, FCN-DK architecture allows n number of input bands. Second, FCN-DK can accept the input image of any dimension because it uses dilated kernel technique, which helps to maintain the output dimension as same as the input. Third, FCN-DK architecture did not require high computational units compared to other FCN architecture like FCN-VGG16 or FCN-VGG19.

2. What is a suitable grid size?

Different grid sizes were explored during the research with FCN architectures, and it was discovered that a 125 x 125 grid size (patch size) works well for slum mapping.

3. Which technique can be used to interpolate the feature vector of SVI into the 2-dimensional (2D) space of RSI?

Initially, 128 feature vectors from each SVI were extracted using fine-tuned Places365-VGG16, and then 128 feature vectors of each SVI were reduced to 32 feature vectors using PCA with a variance of 58.35%. Finally, 32 feature vectors of each SVI are interpolated into a 2D space with the exact spatial resolution (0.4 m) of RSI using the IDW spatial interpolation technique in ArcGIS.

4. How to deal with the incomplete data of SVI?

The IDW approach assumes that the influence of the variable being mapped decreases as the distance from the sampled location increases. IDW is used to estimate the value of unknown points using the points with known values. In this research, the unknown points are those areas where the SVI was unavailable and the known points where the SVI is available.

Some areas were quite far from the available SVI locations, due to which there is a minimal influence of the available SVI in those areas. Therefore, as discussed in Section7.2.2, Modified FCN-DK6 performs the same as FCN-DK6 because of the lack of added ground-level information.

III. To investigate the significance of using SVI for mapping slums

In this research, different FCN architectures were set up using different datasets, i.e., FCN-DK6 used RSI alone, and FCN-DK6-i and Modified FCN-DK6 used a combination of RSI and SVI. Further, FCN architectures results were compared to investigate the significance of using SVI with RSI.

1. What is the added value of combining SVI and RSI for mapping slums?

The features extracted from SVI were integrated with RSI for mapping slums, and SVI features were extracted using fine-tuned Places365-VGG16 network as discussed in Section 5.3.2.3. Thus fine-tuned Places365-VGG16 plays a crucial role in integrating RSI with SVI because if the features extracted from the SVI closely represent slum features, then there is a high probability of getting better accuracy after integrating those features with RSI for slum mapping, discussed in detail in Section 7.3.2 point (II).

Figure 7.7 and Figure 7.8 compare the outcome of FCN-DK6 and Modified FCN-DK6, and it can be concluded that adding SVI with RSI increases the accuracy of the predicted slum map. Thus, ground-level information extracted from SVI helps to understand the urban scene better than using RSI alone. In contrast, how the RSI is integrated with SVI is also important because not every time the integration of RSI with SVI leads to a better result, as shown in Figure 6.8. Thus the result of FCN-DK6-i was slightly less than the result of FCN-DK6.

8.2. Limitations and Recommendations

This research explores integrating two different datasets, i.e., RSI and SVI, for slum mapping using a DL approach. This research put forward some benefits and drawbacks of integrating RSI and SVI for slum mapping. Although our method successfully integrated RSI and SVI and got slightly more accuracy than RSI alone. Still, the research has certain limitations, and further research is required.

Limitations of our work are as follows:

1. The lack of information related to the official slum reference data of 2017, i.e., the exact indicators used to delineate the official slum reference map and how the different categories of slums were differentiated, was not clearly defined by the local government of Jakarta.
2. We do not have the ground knowledge of slums in Jakarta in detail, which made delineating the tweaked slum reference map difficult. However, we have used the local expert's knowledge to understand the context of slums in Jakarta as much as possible.
3. Due to the limited excess to download the GSV, we couldn't generate an overview of the entire study area. If we had a chance to do it for our study area, we could have known how well Modified FCN-DK6 understands slums in our study area. It might be possible that the network would have mapped some slum areas within the non-slum area (others) because there is always a chance of error while generating a slum reference map (tweaked slum reference map) through visual interpretation.
4. While generating the feature maps, we couldn't generate the map of the entire area at once due to the limited computational capacity of the laptop processor. Therefore, we have used model builder in ArcGIS to generating the feature maps, but it was quite time-consuming.

Recommendations for further research are as follows:

1. More ground-level information and ancillary data, such as population density per building footprint/area, access to water and sanitation, and hazardous areas, can be incorporated to delineate the slum reference map more precisely.
2. The comparative study can be carried out for concatenating the SVI with RSI at different convolutional blocks in Modified FCN-DK6.

3. Different interpolation techniques can be explored for generating feature maps from SVI, like kriging.
4. Different FCN architecture can be explored with various fusion methodologies for integrated RSI and SVI.
5. A comparative study can be done between the combination of high-resolution aerial imagery with SVI and VHR satellite imagery with SVI for slum mapping. The motive behind using aerial imagery is that it contains much more detailed information than RSI, such as roads' quality, which might help the FCN network understand the urban scene better when integrated with SVI.
6. A comparative study can be done on the transferability of Modified FCN-DK6 to integrate VHR imagery from different sensors with SVI for slum mapping.
7. Places365-VGG16 network can be fine-tuned more precisely for different categories of slums in the context of Jakarta to understand the ground-level features on which slum categories can be differentiated.
8. Further, Place365-VGG16 and Modified FCN-DK6 can be fine-tuned on the different regions of Jakarta, i.e., by training the networks for varying slums characteristics if they exist within Jakarta. Finally, the slum map can be generated for the entire Jakarta by using fine-tuned Modified FCN-DK6 network.

LIST OF REFERENCES

Ajami, A., Kuffer, M., Persello, C., & Pfeffer, K. (2019). Identifying a slums' degree of deprivation from VHR images using convolutional neural networks. *Remote Sensing*, *11*(11). https://doi.org/10.3390/rs11111282

Alzamil, W. S. (2018). Evaluating Urban Status of Informal Settlements in Indonesia: A Comparative Analysis of Three Case Studies in North Jakarta. *Journal of Sustainable Development*, *11*(4), 148. https://doi.org/10.5539/jsd.v11n4p148

Arimah, B. C. (2011). The Face of Urban Poverty: Explaining the Prevalence of Slums in Developing Countries. *Urbanization and Development: Multidisciplinary Perspectives*. https://doi.org/10.1093/acprof:oso/9780199590148.003.0008

Blaschke, T., Hay, G. J., Kelly, M., Lang, S., Hofmann, P., Addink, E., … Tiede, D. (2014). Geographic Object-Based Image Analysis - Towards a new paradigm. *ISPRS Journal of Photogrammetry and Remote Sensing*, *87*, 180–191. https://doi.org/10.1016/j.isprsjprs.2013.09.014

Brownlee, J. (2018). How to Calculate Principal Component Analysis (PCA) from Scratch in Python. Retrieved July 22, 2021, from https://machinelearningmastery.com/calculate-principal-component-analysis-scratch-python/

Cao, R., Zhu, J., Tu, W., Li, Q., Cao, J., Liu, B., … Qiu, G. (2018). Integrating aerial and street view images for urban land use classification. *Remote Sensing*, *10*(10), 1–23. https://doi.org/10.3390/rs10101553

Central Bureau of Statistics. (2019). Statistical Reference Information System - View Indicator. Retrieved July 30, 2021, from https://sirusa.bps.go.id/sirusa/index.php/indikator/1441

Central Bureau of Statistics DKI Jakarta. (2020). *Indicator Development Purpose Sustainable DKI Jakarta Province*. Retrieved from https://jakarta.bps.go.id/publication/2021/06/28/ec2c87a8b63070211a7bef7e/indikator-tujuan-pembangunan-berkelanjutan-provinsi-dki-jakarta-2020.html

Duque, J. C., Patino, J. E., & Betancourt, A. (2017). Exploring the potential of machine learning for automatic slum identification from VHR imagery. *Remote Sensing*, *9*(9), 1–23. https://doi.org/10.3390/rs9090895

Duque, J. C., Patino, J. E., Ruiz, L. A., & Pardo-Pascual, J. E. (2015). Measuring intra-urban poverty using land cover and texture metrics derived from remote sensing data. *Landscape and Urban Planning*, *135*, 11–21. https://doi.org/10.1016/j.landurbplan.2014.11.009

European Space Agency. (2021). WorldView-2 full archive and tasking - Earth Online. Retrieved July 31, 2021, from https://earth.esa.int/eogateway/catalog/worldview-2-full-archive-and-tasking?text=worldview-2

Gao, Y. (2020). *Assessing the spatial transferability of fully convolutional networks for slum mapping*. Retrieved from https://essay.utwente.nl/84935/

Hoeser, T., & Kuenzer, C. (2020). Object detection and image segmentation with deep learning on Earth observation data: A review-part I: Evolution and recent trends. *Remote Sensing*, *12*(10). https://doi.org/10.3390/rs12101667

Hofmann, P., Strobl, J., Blaschke, T., & Kux, H. (2008). *Detecting informal settlements from QuickBird data in Rio de Janeiro using an object based approach BT - Object-Based Image Analysis: Spatial Concepts for Knowledge-Driven Remote Sensing Applications* (T. Blaschke, S. Lang, & G. J. Hay, Eds.). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-77058-9_29

Ibrahim, M. R., Haworth, J., & Cheng, T. (2019). URBAN-i: From urban scenes to mapping slums, transport modes, and pedestrians in cities using deep learning and computer vision. *Environment and Planning B: Urban Analytics and City Science*, *0*(0), 1–18. https://doi.org/10.1177/2399808319846517

Irawaty, D. T. (2018). Jakarta's Kampungs: Their History and Contested Future. *ProQuest Dissertations and Theses*, 89. Retrieved from https://search.proquest.com/dissertations-theses/jakartas-kampungs-their-history-contested-future/docview/2065034911/se-2?accountid=41849

Joshi, P., Sen, S., & Hobson, J. (2002). Experiences with surveying and mapping Pune and Sangli slums on a geographical information system (GIS). *Environment and Urbanization*, *14*(2), 225–240.

https://doi.org/10.1177/095624780201400218

Ke, Q., Liu, J., Bennamoun, M., An, S., Sohel, F., & Boussaid, F. (2018). Computer vision for human-machine interaction. In *Computer Vision For Assistive Healthcare*. Elsevier Ltd. https://doi.org/10.1016/B978-0-12-813445-0.00005-8

Kohli, D. (2015). Identifying and classifying slum areas using remote sensing. In *Doctoral dissertation*. Retrieved from http://purl.org/utwente/doi/10.3990/1.9789036540087

Kohli, D., Kerle, N., & Sliuzas, R. (2012). Local ontologies for object-based slum identification and classification. *Proceedings of the 4th GEOBIA*, (May), 201. Retrieved from https://www.researchgate.net/profile/Norman_Kerle/publication/230667153_Local_ontologies_for_object-based_slum_identification_and_classification/links/0912f502b543ca854a000000/Local-ontologies-for-object-based-slum-identification-and-classification.pdf%0A

Kohli, D., Warwadekar, P., Kerle, N., Sliuzas, R., & Stein, A. (2013). Transferability of object-oriented image analysis methods for slum identification. *Remote Sensing*, Vol. 5, pp. 4209–4228. https://doi.org/10.3390/rs5094209

Kuffer, M., Pfeffer, K., & Sliuzas, R. (2016). Slums from space-15 years of slum mapping using remote sensing. *Remote Sensing*, *8*(6). https://doi.org/10.3390/rs8060455

Kuffer, M., Pfeffer, K., Sliuzas, R., & Baud, I. (2016). Extraction of Slum Areas From VHR Imagery Using GLCM Variance. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*(5), 1830–1840. https://doi.org/10.1109/JSTARS.2016.2538563

Kuffer, M., Pfeffer, K., Sliuzas, R., Taubenbock, H., Baud, I., & Van Maarseveen, M. (2018). Capturing the Urban Divide in Nighttime Light Images from the International Space Station. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *11*(8), 2578–2586. https://doi.org/10.1109/JSTARS.2018.2828340

Legarias, T. M., Nurhasana, R., & Irwansyah, E. (2020). Building Density Level of Urban Slum Area in Jakarta. *Geosfera Indonesia*, *5*(2), 268. https://doi.org/10.19184/geosi.v5i2.18547

Leonita, G., Kuffer, M., Sliuzas, R., & Persello, C. (2018). Machine learning-based slum mapping in support of slum upgrading programs: The case of Bandung City, Indonesia. *Remote Sensing*, *10*(10). https://doi.org/10.3390/rs10101522

Lilford, R., Kyobutungi, C., Ndugwa, R., Sartori, J., Watson, S. I., Sliuzas, R., … Ezeh, A. (2019). Because space matters: Conceptual framework to help distinguish slum from non-slum urban areas. *BMJ Global Health*, *4*(2). https://doi.org/10.1136/bmjgh-2018-001267

Liu, Qinghui & Salberg, Arnt-Børre and Jenssen, R. (2018). *A Comparison of Deep Learning Architectures for Semantic Mapping of Very High Resolution Images*. Retrieved from https://doi.org/10.1109/IGARSS.2018.8518533

Liu, L., Zhenwei, S., Pan, B., Zhang, N., Luo, H., & Lan, X. (2020). *Multiscale Deep Spatial Feature Extraction Using Virtual RGB Image for Hyperspectral Imagery Classification*. Retrieved from https://doi.org/10.3390/rs12020280

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation Jonathan. *Intas Polivet*, *10*(2), 227–228. Retrieved from https://doi.org/10.1109/CVPR.2015.7298965

Lundström, D. (2017). *Data-efficient Transfer Learning with Pre-trained Networks*. Retrieved from https://liu.diva-portal.org/smash/get/diva2:1112122/FULLTEXT01.pdf

Mahabir, R., Croitoru, A., Crooks, A., Agouris, P., & Stefanidis, A. (2018). A Critical Review of High and Very High-Resolution Remote Sensing Approaches for Detecting and Mapping Slums: Trends, Challenges and Emerging Opportunities. *Urban Science*, *2*(1), 8. https://doi.org/10.3390/urbansci2010008

Martinez, R., & Nurlina Masron, I. (2020). *Jakarta A city of cities*. Retrieved from https://doi.org/10.1016/j.cities.2020.102868

Mboga, N., Persello, C., Bergado, J. R., & Stein, A. (2017). Detection of informal settlements from VHR images using convolutional neural networks. *Remote Sensing*, *9*(11). https://doi.org/10.3390/rs9111106

Michael, X., Neal, J., Burke, M., Lobell, D., & Ermon, S. (2016). *Transfer Learning from Deep Features for Remote Sensing and Poverty Mapping*. Retrieved from https://arxiv.org/abs/1510.00098

Minister of National Development Planning [MNDP] Indonesia. (2019). Voluntary National Reviews Database. *Sustainable Development Knowledge Platform*. Retrieved from https://sustainabledevelopment.un.org/vnrs/

Mohammad, H., & Sulaiman. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 01–11.

https://doi.org/10.5121/ijdkp.2015.5201

Nijman, J. (2008). Against the odds: Slum rehabilitation in neoliberal Mumbai. *Cities*, *25*(2), 73–85. https://doi.org/10.1016/j.cities.2008.01.003

Nurdiansyah, A. (2018). Urban Slum Upgrading Policy In Jakarta (Case Study: Kampung Deret Program Implementation). *The Indonesian Journal of Planning and Development*, *3*(1), 19. https://doi.org/10.14710/ijpd.3.1.19-31

Pangeran, A., & Akbar, R. J. (2020). *Cities for Marginal Community: Lesson Learned from Indonesia's Slum Alleviation Program*. Retrieved from https://www.isocarp-institute.org/wp-content/uploads/2020/08/Review15_Cities-for-marginal-community.pdf

Persello, C., & Stein, A. (2017). Deep Fully Convolutional Networks for the Detection of Informal Settlements in VHR Images. *IEEE Geoscience and Remote Sensing Letters*, *14*(12), 2325–2329. https://doi.org/10.1109/LGRS.2017.2763738

Pratomo, J., Kuffer, M., Kohli, D., & Martinez, J. (2018). Application of the trajectory error matrix for assessing the temporal transferability of OBIA for slum detection. *European Journal of Remote Sensing*, *51*(1), 838–849. https://doi.org/10.1080/22797254.2018.1496798

Pratomo, J., Kuffer, M., Martinez, J., & Kohli, D. (2017). Coupling uncertainties with accuracy assessment in object-based slum detections, case study: Jakarta, Indonesia. *Remote Sensing*, *9*(11). https://doi.org/10.3390/rs9111164

Putranto, S. (2009). *Redefining the Spatial Form of Urban Village in Mega Kuningan Jakarta as A New Urban Integrator: A Study of Socio-economic Aspect in the Forming of Urban Spatial Configuration*. (September). Retrieved from http://hub.hku.hk/handle/10722/56857

Rahman, M. A., & Wang, Y. (2016). *Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation* (G. Bebis, R. Boyle, B. Parvin, D. Koracin, F. Porikli, S. Skaff, … T. Isenberg, Eds.). Cham: Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-319-50835-1 22

Rukmana, D. (2008). Planning the Megacity: Jakarta in the Twentieth Century. *Journal of the American Planning Association*, *74*(2), 263–264. https://doi.org/10.1080/01944360801940995

Sergey Ioffe and Christian Szegedy. (2015). *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. Retrieved from http://proceedings.mlr.press/v37/ioffe15.html

Stark, T., Wurm, M., Taubenbock, H., & Zhu, X. X. (2019). Slum mapping in imbalanced remote sensing datasets using transfer learned deep features. *2019 Joint Urban Remote Sensing Event, JURSE 2019*, 1–4. https://doi.org/10.1109/JURSE.2019.8808965

Taubenböck, H., & Kraff, N. J. (2014). The physical face of slums A structural comparison of slums in Mumbai, India based on remotely sensed data. *DLR Deutsches Zentrum Fur Luft- Und Raumfahrt e.V. - Forschungsberichte*, Vol. 2019-Janua, pp. 413–442. Retrieved from https://doi.org/10.1007/s10901-013-9333-x

Teerapong, Panboonyuen Kulsawasd, J., Siam, L., Panu, S., & Peerapon, V. (2017). *Road segmentation of remotely-sensed images using deep convolutional neural networks with landscape metrics and conditional random fields*. Retrieved from https://doi.org/10.3390/rs9070680

The World Bank. (2016). *National Slum Upgrading Project*. Retrieved from https://monitoring.skp-ham.org/wp-content/uploads/2020/03/Proposed-Loan-KOTAKU-June-9-2016.pdf

United Nations-Department of Economic and Social Affairs [UNDESA]. (2015). THE 17 GOALS | Sustainable Development. Retrieved June 19, 2021, from https://sdgs.un.org/goals

United Nations-Department of Economic and Social Affairs [UNDESA]. (2017). Global indicator framework for the Sustainable Development Goals and targets of the 2030 Agenda for Sustainable Development. *Work of the Statistical Commission Pertaining to the 2030 Agenda for Sustainable Development*, 1–21. Retrieved from https://unstats.un.org/sdgs/indicators/Global Indicator Framework after refinement_Eng.pdf

United Nations-Department of Economic and Social Affairs [UNDESA]. (2018, May 16). World population projected to live in urban areas. Retrieved November 8, 2020, from https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html

United Nations-Habitat [UN-Habitat]. (2003). THE CHALLENGE OF SLUMS. In *United Nations Human Settlements Programme* (Vol. 238). https://doi.org/10.1006/abio.1996.0254

United Nations-Habitat [UN-Habitat]. (2013). *Report of the Special Rapporteur on adequate housing, Indoniesia*. Retrieved from https://www.ohchr.org/EN/HRBodies/HRC/RegularSessions/Session25/Documents/A-HRC-

25-54-Add1_en.doc.

United Nations-Habitat [UN-Habitat]. (2018). *Pro-Poor Climate Action in Informal Settlements*. Retrieved from https://unhabitat.org/pro-poor-climate-action-in-informal-settlement

United Nations Department of Economic and Social Affairs [UNDESA]. (2020). SDG Indicators. Retrieved November 8, 2020, from https://unstats.un.org/sdgs/report/2020/goal-11/

United Nations Development Programme [UNDP]. (2016). *From the MDGs to Sustainable Development for All*. Retrieved from https://reliefweb.int/report/world/mdgs-sustainable-development-all-lessons-15-years-practice

Verma, D., Jana, A., & Ramamritham, K. (2019). Transfer learning approach to map urban slums using high and medium resolution satellite imagery. *Habitat International*, *88*(April), 101981. https://doi.org/10.1016/j.habitatint.2019.04.008

Workman, Scott, Zhai, M., Crandall, D. J., & Jacobs, N. (2017). *A Unified Model for Near and Remote Sensing*. (1), 8–11. Retrieved from https://doi.org/10.1109/ICCV.2017.293

Workman, ScottJabocs, Zhai, M., Crandall, D. J., & Jacobs, N. (2017). A Unified Model for Near and Remote Sensing. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-Octob*(1), 2707–2716. https://doi.org/10.1109/ICCV.2017.293

Wurm, M., Stark, T., Zhu, X. X., Weigand, M., & Taubenböck, H. (2019). Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, *150*(May 2018), 59–69. https://doi.org/10.1016/j.isprsjprs.2019.02.006

Xia, X., Koeva, M., & Persello, C. (2019). Extracting Cadastral Boundaries from UAV Images Using Fully Convolutional Networks. *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2455–2458. https://doi.org/10.1109/IGARSS.2019.8898156

Zahidi, I., Yusuf, B., Hamedianfar, A., Shafri, H. Z. M., & Mohamed, T. A. (2015). Object-based classification of QuickBird image and low point density LIDAR for tropical trees and shrubs mapping. *European Journal of Remote Sensing*, *48*, 423–446. https://doi.org/10.5721/EuJRS20154824

Zhao, W., Bo, Y., Chen, J., Tiede, D., Thomas, B., & Emery, W. J. (2019). Exploring semantic elements for urban scene recognition: Deep integration of high-resolution imagery and OpenStreetMap (OSM). *ISPRS Journal of Photogrammetry and Remote Sensing*, *151*(March), 237–250. https://doi.org/10.1016/j.isprsjprs.2019.03.019

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009

Zhu, J. (2010). Symmetric Development of Informal Settlements and Gated Communities: Capacity of the State - The Case of Jakarta, Indonesia. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1716585

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). *Deep learning in remote sensing: a review*. (December). https://doi.org/10.1109/MGRS.2017.2762307

APPENDICES

Annexure-I: Detailed Model Description

FCN-DK6 Architecture

Dks are used to keep the output images with the exact dimension and resolution of the input image without the deconvolutional layer (Persello and Stein, 2017). In each convolutional block, DK inserts zeros to input before feeding it to the convolutional layer and after the output of the LeakyReLU layer to maintain the size of output as same as the input. The number of zeroes to be inserted for each convolutional block can be calculated by d-1, where d is the dilation factor. After completing each convolutional block, the output kernel dimensions can be calculated using equation (i), where H is kernel height and W is kernel width. Thus, the receptive field increased exponentially with each convolutional block without increasing the count of learnable parameters. As shown in Figure 1, the receptive field of d = 1 and d = 2.

$$H' \times W' = [d \times (H-1) + 1] \times [d \times (W-1) + \ 1] \qquad \text{(i)}$$

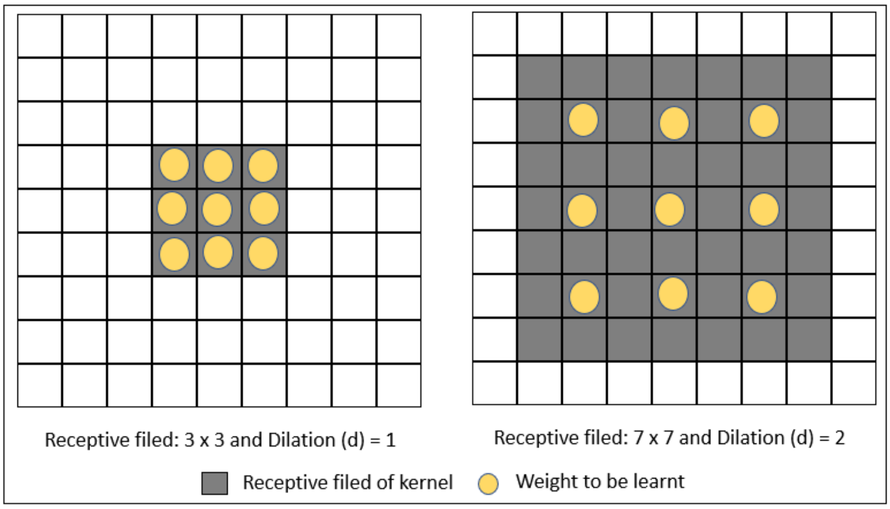

Receptive filed: 3 x 3 and Dilation (d) = 1
Receptive filed: 7 x 7 and Dilation (d) = 2
Receptive filed of kernel
Weight to be learnt

Figure 1: Shows receptive filed

Persello and Stein (2017) introduce FCN-DK architecture such as FCN-DK6 with six convolutional blocks adopted for this research. Each convolutional block consists of six layers: one zero-padding, one convolutional, one batch normalization, one leaky Rectified Linear Units (lReLU), and one max pooling. After all six blocks, one dropout layer, one classification layer are present. Different layers in the convolutional block have different functions: (i) The zero-padding layer is used to keep the dimension of the output image as same as the input image, which makes FCN-DK architecture quite flexible with the input image dimensions (Persello & Stein, 2017). (ii) The convolutional layer learns the features from the input. (iii) The batch normalization layer normalizes the mini-batch input to avoid the internal variation shift problem while training the model. Hence, the input layer's distribution will be changed accordingly with the change of the learnable parameter of the previous layer (Sergey Ioffe and Christian Szegedy, 2015). (iv) IReLU is called an activation function in the network, identifying whether the pixel belongs to slum or non-slum class (Xia, Koeva, & Persello, 2019). (v) The max-pooling layer is used to reduce the dimensionality of the input layer. Hence, it reduces the number of training parameters, computational cost and restricts the model's overfitting problem. (vi) The dropout layer is also used to prevent the overfitting of the model. (vii) the classification layer consists of an activation function (softmax) that will help to classify the input into desired output classes. Figure 2 shows the detailed architecture of FCN-DK6 used for this research, and Table 1 presents the structure of FCN-DK6.

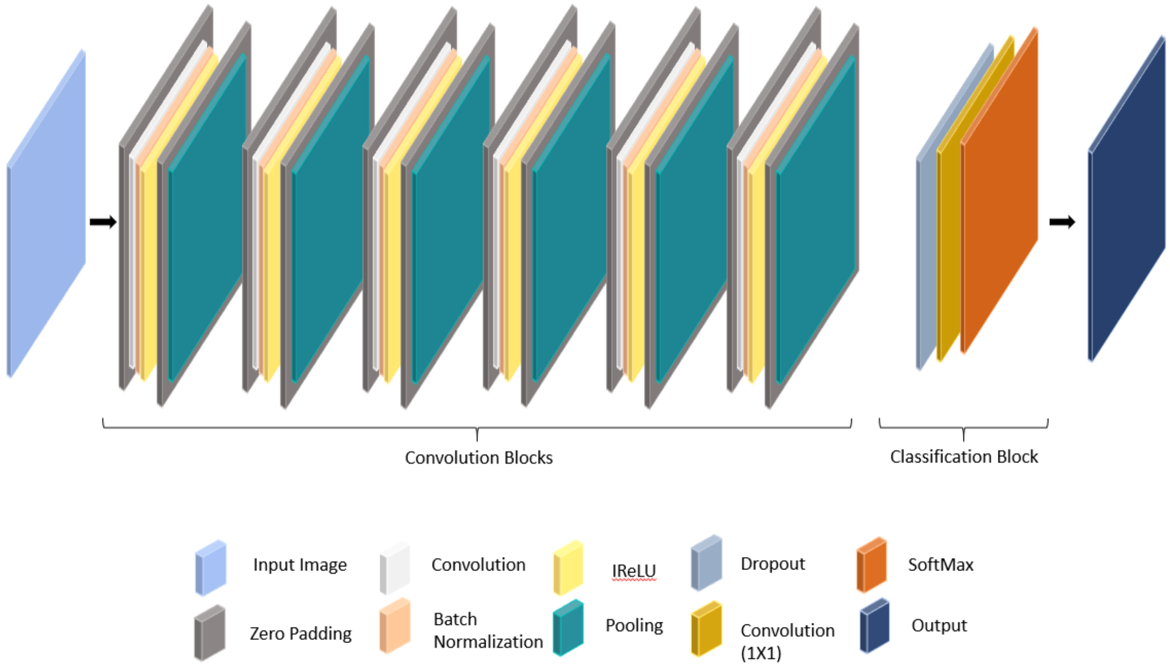


Figure 2: The detailed architecture of FCN-DK6 used for this research

Table 1: Present the structure of FCN-DK6

Block	Layer	Hyper-Parameter
1	ZeroPadding2D	Padding: 2x2
	Convolutional2D	Number of Filters: 16 Kernel Size: 5x5 Dilation Rate: 1x1
	Batch Normalization	Axis: 3
	leaky Rectified Linear Units	Alpha: 0.1
	ZeroPadding2D	Padding: 2x2
	MaxPooling2D	Pool Size: 5x5 Strides: 1x1
2	ZeroPadding2D	Padding: 4x4
	Convolutional2D	Number of Filters: 32 Kernel Size: 5x5 Dilation Rate: 2x2
	Batch Normalization	Axis: 3
	leaky Rectified Linear Units	Alpha: 0.1
	ZeroPadding2D	Padding: 4x4
	MaxPooling2D	Pool Size: 9x9 Strides: 1x1
3	ZeroPadding2D	Padding: 6x6

	Convolutional2D	Number of Filters: 32 Kernel Size: 5x5 Dilation Rate: 3x3
	Batch Normalization	Axis: 3
	leaky Rectified Linear Units	Alpha: 0.1
	ZeroPadding2D	Padding: 6x6
	MaxPooling2D	Pool Size: 13x13 Strides: 1x1
4	ZeroPadding2D	Padding: 8x8
	Convolutional2D	Number of Filters: 32 Kernel Size: 5x5 Dilation Rate: 4x4
	Batch Normalization	Axis: 3
	leaky Rectified Linear Units	Alpha: 0.1
	ZeroPadding2D	Padding: 8x8
	MaxPooling2D	Pool Size: 17x17 Strides: 1x1
5	ZeroPadding2D	Padding: 10x10
	Convolutional2D	Number of Filters: 32 Kernel Size: 5x5 Dilation Rate: 5x5
	Batch Normalization	Axis: 3
	leaky Rectified Linear Units	Alpha: 0.1
	ZeroPadding2D	Padding: 10x10
	MaxPooling2D	Pool Size: 21x21 Strides: 1x1
6	ZeroPadding2D	Padding: 12x12
	Convolutional2D	Number of Filters: 32 Kernel Size: 5x5 Dilation Rate: 6x6
	Batch Normalization	Axis: 3
	leaky Rectified Linear Units	Alpha: 0.1
	ZeroPadding2D	Padding: 12x12
	MaxPooling2D	Pool Size: 25x25 Strides: 1x1
Classification	Dropout	Rate: 0.25
	Convolutional	Number of Filters: 2 (number of classes) Kernel Size: 1x1
	Activation	SoftMax

VGG16 Architecture

VGG stands for Visual Geometry Group developed at Oxford University. Two factors make the VGG model simple for use. First, the network's extensive use of 3x3 convolutions, and second, the number of feature maps is doubled after the max-pooling layer of 2x2 with stride 2, i.e., this arrangement eliminates the need for tuning convolutional filter sizes and individual layer sizes (Lundström, 2017).

The VGG 16 architecture proposed for this research consists of 5 convolutional blocks and one classification block. The first two convolutional blocks consist of three layers: two convolutional layers and one max-pooling layer. The last three convolutional blocks consist of four layers: three convolutional layers and one max-pooling layer. The convolutional block's output is pass-through from the global max-pooling layer, which further connects to the classification block. The classification block consists of five layers: one flatten layer, two dense layers, one dropout layer, and one classification layer. All of the layers mentioned above have different functions: (i) The convolutional layer is the key layer of the network. The convolutional layer consists of a different set of filters responsible for extracting features from the input image, and each filter will produce a feature map as output (Ke et al., 2018). (ii) The max-pooling layer is used to minimize the spatial dimensionality of the feature map and control the overfitting of the model, as shown in Figure5.9. (iii) The global max-pooling layer is used to compress the whole image more efficiently using strides (stride is used to move the filter over the width and height), as shown in Figure 3. (iv) The flatten layer is used to flatten the input. (v) The dense layer is provided along with ReLU activation. The dense layer is used to change the dimension of the input. (vi) The dropout layer is also used to prevent the overfitting of the model. (vii) the classification layer is used to classify the input into desired out classes. Figure 4 shows the detailed architecture of VGG16 used for this research, and Table 2 presents the structure of VGG16.

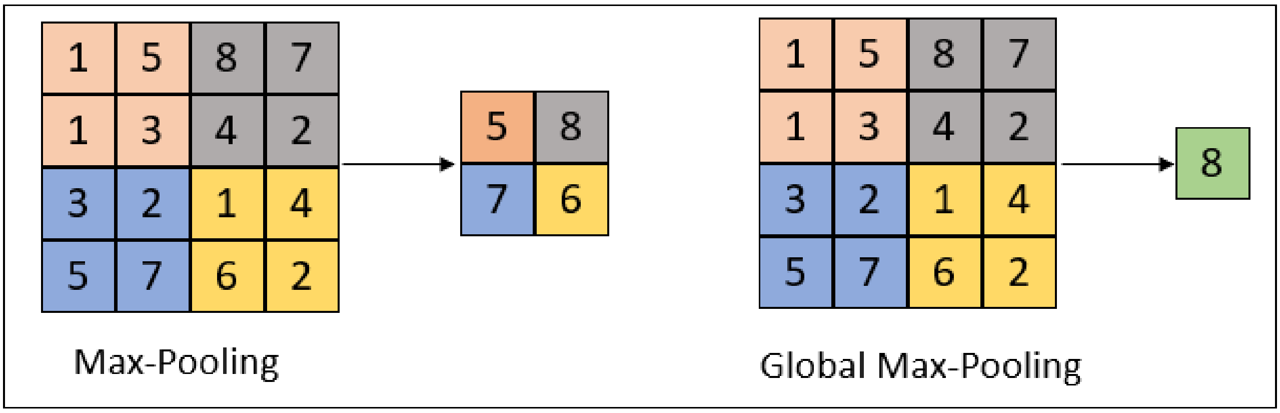


Figure 3: Show the Max pooling and Global pooling functions

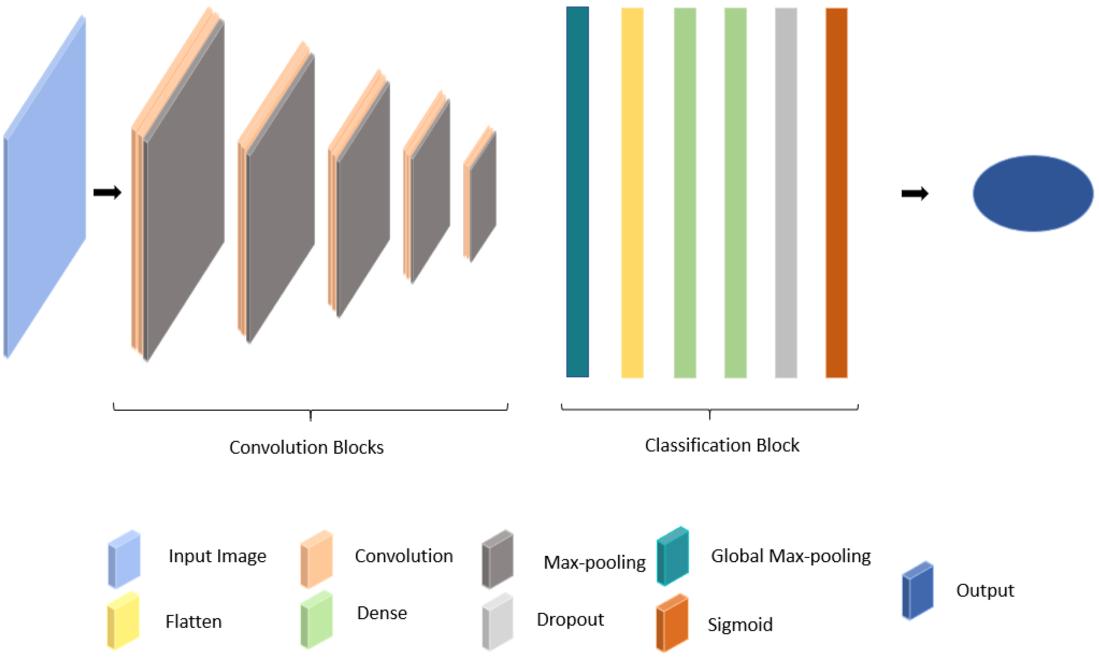


Figure 4: Show the detailed architecture of Places365-VGG16 used for this research

Table 2: Presents the structure of Places365-VGG16

Block	Layer	Hyper-Parameter
1	Convolutional2D	Number of Filters: 64 Kernel Size: 3x3 Stride: 1x1 Kernel Regularizer: l2 Padding: 'same' Activation: 'relu'
	Convolutional2D	Number of Filters: 64 Kernel Size: 3x3 Stride: 1x1 Kernel Regularizer: l2 Padding: 'same' Activation: 'relu'
	MaxPooling2D	Pool Size: 2x2 Strides: 2x2 Padding: 'valid'
2	Convolutional2D	Number of Filters: 128 Kernel Size: 3x3 Stride: 1x1 Kernel Regularizer: l2 Padding: 'same'

		Activation: 'relu'
	Convolutional2D	Number of Filters: 128 Kernel Size: 3x3 Stride: 1x1 Kernel Regularizer: l2 Padding: 'same' Activation: 'relu'
	MaxPooling2D	Pool Size: 2x2 Strides: 2x2 Padding: 'valid'
3	Convolutional2D	Number of Filters: 256 Kernel Size: 3x3 Stride: 1x1 Kernel Regularizer: l2 Padding: 'same' Activation: 'relu'
	Convolutional2D	Number of Filters: 256 Kernel Size: 3x3 Stride: 1x1 Kernel Regularizer: l2 Padding: 'same' Activation: 'relu'
	Convolutional2D	Number of Filters: 256 Kernel Size: 3x3 Stride: 1x1 Kernel Regularizer: l2 Padding: 'same' Activation: 'relu'
	MaxPooling2D	Pool Size: 2x2 Strides: 2x2 Padding: 'valid'
4	Convolutional2D	Number of Filters: 512 Kernel Size: 3x3 Stride: 1x1 Kernel Regularizer: l2 Padding: 'same' Activation: 'relu'
	Convolutional2D	Number of Filters: 512 Kernel Size: 3x3 Stride: 1x1 Kernel Regularizer: l2 Padding: 'same' Activation: 'relu'
	Convolutional2D	Number of Filters: 512 Kernel Size: 3x3 Stride: 1x1 Kernel Regularizer: l2 Padding: 'same' Activation: 'relu'
	MaxPooling2D	Pool Size: 2x2

		Strides: 2x2 Padding: 'valid'
5	Convolutional2D	Number of Filters: 512 Kernel Size: 3x3 Stride: 1x1 Kernel Regularizer: l2 Padding: 'same' Activation: 'relu'
	Convolutional2D	Number of Filters: 512 Kernel Size: 3x3 Stride: 1x1 Kernel Regularizer: l2 Padding: 'same' Activation: 'relu'
	Convolutional2D	Number of Filters: 512 Kernel Size: 3x3 Stride: 1x1 Kernel Regularizer: l2 Padding: 'same' Activation: 'relu'
	MaxPooling2D	Pool Size: 2x2 Strides: 2x2 Padding: 'valid'
	GlobalMaxPooling2D	
Classification	Flatten	
	Dense	Units: 512 Activation: 'relu'
	Dense	Units: 128 Activation: 'relu'
	Dropout	Rate: 0.5
	Dense	Unit: 2 (number of classes) Activation: 'sigmoid'

Modified FCN-DK6 Architecture

Table 4 presents the structure of Modified FCN-DK6.

Table 4: Presents the structure of Modified FCN-DK6

Block	Layer	Hyper-Parameter
1	ZeroPadding2D	Padding: 2x2
	Convolutional2D	Number of Filters: 16 Kernel Size: 5x5 Dilation Rate: 1x1
	Batch Normalization	Axis: 3
	leaky Rectified Linear Units	Alpha: 0.1
	ZeroPadding2D	Padding: 2x2
	MaxPooling2D	Pool Size: 5x5 Strides: 1x1
2	ZeroPadding2D	Padding: 4x4
	Convolutional2D	Number of Filters: 32 Kernel Size: 5x5 Dilation Rate: 2x2
	Batch Normalization	Axis: 3
	leaky Rectified Linear Units	Alpha: 0.1
	ZeroPadding2D	Padding: 4x4
	MaxPooling2D	Pool Size: 9x9 Strides: 1x1
	Concatenate	
3	ZeroPadding2D	Padding: 6x6
	Convolutional2D	Number of Filters: 32 Kernel Size: 5x5 Dilation Rate: 3x3
	Batch Normalization	Axis: 3
	leaky Rectified Linear Units	Alpha: 0.1
	ZeroPadding2D	Padding: 6x6
	MaxPooling2D	Pool Size: 13x13 Strides: 1x1
4	ZeroPadding2D	Padding: 8x8
	Convolutional2D	Number of Filters: 32 Kernel Size: 5x5 Dilation Rate: 4x4
	Batch Normalization	Axis: 3
	leaky Rectified Linear Units	Alpha: 0.1
	ZeroPadding2D	Padding: 8x8
	MaxPooling2D	Pool Size: 17x17 Strides: 1x1
5	ZeroPadding2D	Padding: 10x10
	Convolutional2D	Number of Filters: 32 Kernel Size: 5x5

		Dilation Rate: 5x5
	Batch Normalization	Axis: 3
	leaky Rectified Linear Units	Alpha: 0.1
	ZeroPadding2D	Padding: 10x10
	MaxPooling2D	Pool Size: 21x21 Strides: 1x1
6	ZeroPadding2D	Padding: 12x12
	Convolutional2D	Number of Filters: 32 Kernel Size: 5x5 Dilation Rate: 6x6
	Batch Normalization	Axis: 3
	leaky Rectified Linear Units	Alpha: 0.1
	ZeroPadding2D	Padding: 12x12
	MaxPooling2D	Pool Size: 25x25 Strides: 1x1
Classification	Dropout	Rate: 0.25
	Convolutional	Number of Filters: 2 (number of classes) Kernel Size: 1x1
	Activation	SoftMax