NICOLAUS COPERNICUS UNIVERSITY IN TORUŃ

Bulletin of Geography. Socio-economic Series, No. 58 (2022): 141-154
http://doi.org/10.12775/bgss-2022-0039

BULLETIN OF GEOGRAPHY. SOCIO–ECONOMIC SERIES

journal homepages:
https://apcz.umk.pl/BGSS/index
https://www.bulletinofgeography.umk.pl/

Accuracy evaluation of convolutional neural network classification algorithms for building identification in rural and urban areas from very-high-resolution satellite imagery in Jambi, Indonesia

Daniel Adi Nugroho[1, CFMR], Muhammad Dimyati[2, CFM], Laswanto[3, DFP]

[1,2]*Universitas Indonesia*, Faculty of Mathematics and Natural Sciences, Department of Geography, Gedung H, Kampus UI Depok, Kota Depok, Jawa Barat 16424, Indonesia; e-mail: [1]daniel.adi91@ui.ac.id (*corresponding author*); [2]m.dimyati@sci.ui.ac.id; [3]Department of Public Works and Spatial Planning, Jambi Municipality Government, Indonesia

How to cite:

Nugroho, D.A., Dimyati, M. & Laswanto (2022). Accuracy evaluation of convolutional neural network classification algorithms for building identification in rural and urban areas from very-high-resolution satellite imagery in Jambi, Indonesia. *Bulletin of Geography. Socio-economic Series,* 58(58): 141-154. DOI: http://doi.org/10.12775/bgss-2022-0039

Abstract. Accurate land cover data are essential to a reliable decision-making process; therefore, researchers have turned to novel land cover classification algorithms employing machine learning on high-resolution satellite imagery to improve classification accuracy. The experiment presented in this paper aims to assess the accuracy performance of three patch-based, convolutional neural network architectures (LeNet, VGGNet, and XCeption) in classifying building footprints in rural and urban areas from satellite imagery data, with conventional, pixel-based classification algorithms as a benchmark. The experiment concluded that the CNN classification algorithms consistently outperformed pixel-based algorithms in the accuracy of the resulting building-footprint classification raster. It was also demonstrated that larger image patch size does not always improve classification accuracy in all CNN architectures. This study also revealed that the XCeption architecture performed best among the three CNN architectures assessed, with a 72-pixel patch size having the best accuracy.

Article details:
Received: 15 February 2022
Revised: 14 July 2022
Accepted: 30 October 2022

Key words:
convolutional neural network,
land cover,
building classification,
satellite imagery

Contents:

1. Introduction

Urban planning requires an accurate assessment of the latest real-world condition, including the spatial distribution pattern of the buildings in urban areas, as buildings are the primary sites for housing and production. One of the most cost-effective ways of getting this information is by deriving building-footprint information from satellite imagery (Ayala et al., 2021). At 30-cm resolution, Maxar's Worldview-3 and Worldview-4 imagery have the highest resolution available to date, as compared to the previous generation of commercial high-resolution satellite imagery (Zhu et al., 2020), enabling public access of periodic earth observation with unprecedented detail. However, in December 2018, the WorldView-4 satellite experienced a failure in its control moment gyros, preventing it from collecting imagery (Maxar, 2020), leaving only WorldView-3 as the highest resolution commercial earth observation satellite currently available from Maxar.

In the past, changes in building distribution and density in urban areas were typically monitored by manually interpreting aerial photographs captured on an analogue recording medium. However, with the availability of digital image data from remote-sensing satellites, advancement in classification algorithms and robust off-the-shelf software packages, this task can be carried out more efficiently and accurately. Early image classification algorithms employ statistical methods that tend to be less accurate; however, this has shifted into machine learning methods that enable better predictions through rigorous training processes (Aroma & Raimond, 2016). In this study, buildings are defined as structures with a roof and walls that enable people to live or work inside, such as houses, schools, stores or factories, and do not include built structures without roofs and walls, such as pavements, roads, parking lots or bridges.

The most commonly used machine learning classification algorithms for remotely sensed image data include Random Forest, K-Nearest Neighbour, and Support Vector Machine (Thanh Noi & Kappas, 2017; Shah et al., 2018). The advent of classification algorithms employing neural network architectures has dramatically influenced the accuracy of supervised classification (Kattenborn et al., 2019). Even though such classification methods come with a high cost in computing resources, they have helped increase satellite imagery classification accuracy (Zhang et al., 2018). Previous studies have asserted that Convolutional Neural Network (CNN) architecture significantly outperforms other classification algorithms in satellite image classification (Hu et al., 2018). Another study by Pan et al. (2020) has suggested that built-up area classification from satellite imagery (including the identification and classification of building footprints) is a task that can be reliably solved using CNN (Chawda et al., 2018; Luo et al., 2021).

Many CNN architectures are available in various platforms and frameworks, ranging from classic models such as LeNet and AlexNet to the most advanced ones such as InceptionNet, VGGNet, XCeption and DenseNet (Sultana et al., 2018; Khan et al., 2020). The convolutional neural network architectures that will be studied in this paper are LeNet, VGGNet and XCeption. LeNet was the precursor of modern neural network architecture introduced by Lecun et al. in 1998, and it is included due to its historical significance and relative simplicity. Originally designed to recognize handwritten numeric characters, its architecture is relatively less complicated (having only five convolutional layers) than the other deep CNN architectures; hence, it is faster to train and process, especially when run on today's hardware. The VGGNet devised by Simonyan & Zisserman in 2018 has 26 convolutional layers, and the XCeption architecture (a much deeper CNN architecture designed by Chollet in 2014) has 126 layers. Deep CNN architectures such as VGGNet and XCeption can be used for satellite image classification, facial recognition, and scene and feature detection in multimedia data (Muhammad et al., 2018; Pal et al., 2020; Xiao et al., 2020; Fatima et al., 2021), and both VGGNet and XCeption are currently considered state-of-the-art (Gikunda & Jouandeau, 2019). As the data input, he CNN algorithms require an image patch, whose size depends on the specific network architecture being used. Previous research has demonstrated that a larger image patch size in CNN classifiers can improve classification accuracy (Hamwood et al., 2018).

This study produces raster datasets signifying building footprints derived from the classification process at 1.2-metre resolution, with each pixel carrying a binary attribute (building or non-building). This resolution exceeds the resolution requirement for 1:10,000 mapping (Tobler, 1987; Li et al., 2019). Based on the Indonesian Government Decree Concerning the Map Accuracy for Spatial Planning (Government of Indonesia, 2013), 1:10,000 scale maps are required to perform city-wide spatial planning in Indonesia. By utilizing this raster dataset, this study aims to accomplish a comparative assessment of LeNet, VGGNet, and XCeption network architectures in supervised building classification compared against conventional, pixel-based classification algorithms, such as Random Forest (RF) and k-Nearest Neighbour (KNN), using test points and pixels derived from manually-digitized building polygons as the baseline benchmark for the classification results. RF and KNN were chosen as the comparative baseline since these algorithms are considered to be the state-of-the-art classifier with high overall accuracy (Thanh Noi & Kappas, 2017; Pacheco et al., 2021). Furthermore, we hypothesized that the classification algorithms employing CNN architectures would provide higher classification accuracy compared to RF and KNN, with the larger image patch size and more advanced architecture yielding more accurate results, and we will evaluate the comparative accuracy between various image patch sizes used in the classification training and inference process.

2. Research materials and methods

The overall workflow of this experimental study is illustrated in Figure 2. The study area is located within the vicinity of Jambi City in Jambi Province, Indonesia, around the Batanghari River banks. The large bulk of the built-up area in Jambi City is located on the south riverbank. Jambi City is a rapidly developing city (Hardiani & Lubis, 2017) that is prone to flooding hazards (Fitri & Sumunar, 2019); therefore, rapid and accurate detection of buildings developed along the banks of the Batanghari River is crucial for disaster mitigation efforts.

This experimental study uses a pan-sharpened Worldview-4 imagery mosaic with 30-cm resolution over Jambi City, Indonesia. This mosaic consists of two Worldview-4 scenes: One scene was acquired on 8 August 2017, and the other on 28 March 2018. Both image scenes were radiometrically and geometrically corrected and then combined into a seamless mosaic in December 2018 by the Regional Planning Board of the Jambi Municipality Government. A total of 23 ground control points and SRTM Digital Elevation Model at one arc-second (30-metre) resolution were used to perform geometric correction on the satellite imagery. The satellite image orthomosaic is provided as three-channel natural colour imagery with a bit-depth of 16-bit for each channel (RGB). The near-infrared channel of Worldview-4 imagery was not available for this study. Five distinct training areas were specified throughout the city; each training site is located well outside the test area and is expected to contain standard features typically found within Jambi City. The test area, along with the corresponding WorldView-4 scene coverage, is shown in Figure 1, while the layout of the training area within the study area is shown in Figure 3.

This experiment utilized LeNet, VGGNet and XCeption CNN architectures in Python through

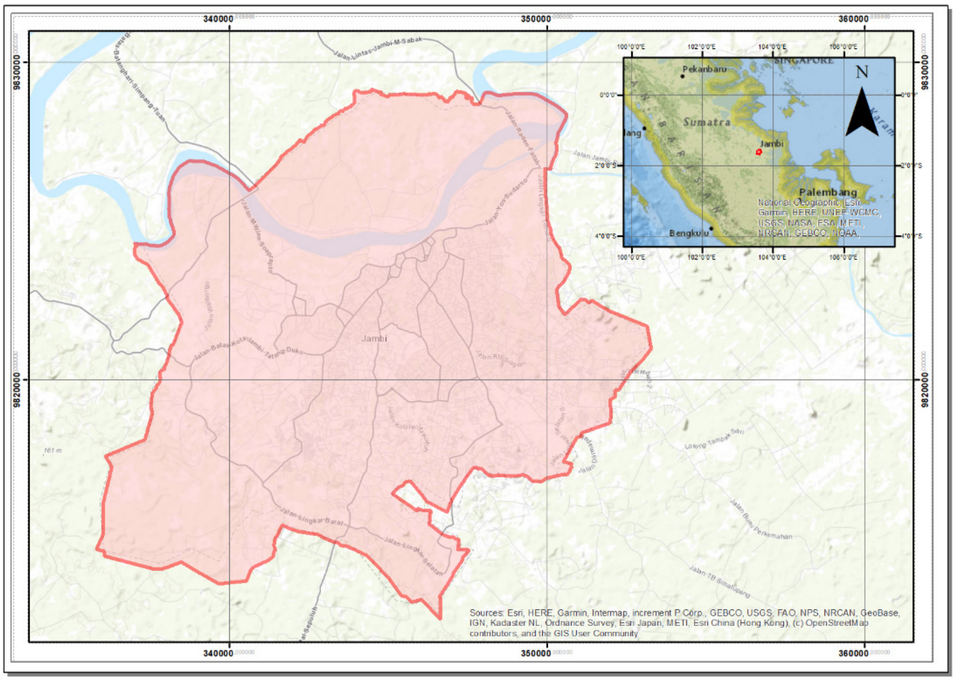
340000
350000
360000
9830000
9820000
Jalan-Lintas-Jambi-M-Sabak
Batas-Batanghari-Simpang-Tuan
Jalan-Raden-Fatah
Jalan-Yos-Sudarso
Jalan-M-Yamin-Soeprapto
Jalan-Batas-Kota-Jambi-Telang-Duku
Jambi
Jalan RB Siagian
Jalan-Kolonel-Abunjani
Jalan-Lingkar-Barat
Jalan-Lingkar-Selatan
Lorong Tambak Sari
Jalan Bumi Perkemahan
Jalan TB Simatupang
161 m
100°0'0"E
102°0'0"E
104°0'0"E
106°0'0"E
0°0'0"
2°0'0"S
4°0'0"S
SINGAPORE
Pekanbaru
Sumatra
Jambi
Palembang
Bengkulu
N
National Geographic, Esri, Garmin, HERE, UNEP-WCMC, USGS, NASA, ESA, METI, NRCAN, GEBCO, NOAA,
Sources: Esri, HERE, Garmin, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), (c) OpenStreetMap contributors, and the GIS User Community

Fig. 1. Red polygon denotes extent of WorldView-4 satellite imagery in study area
Source: own elaboration

TensorFlow and Keras framework, while the KNN and RF algorithms were executed in R language utilizing the Caret package. The computing resource being used in this study is an Intel Core i5-7600K CPU clocked at standard speed (3.8 GHz) with 32 GB of random access memory, while the GPU is an NVIDIA GTX 1070. The VGG-19 architecture was designed with the input image size of 224×224 pixels with a minimum possible input image size of 32×32 pixels, while the XCeption architecture was designed with an input image size of 299×299 pixels with a minimum possible input image size of 71×71 pixels (Chollet, 2015). In comparison, LeNet-5 requires a minimum patch size of 32×32 pixels (Sultana et al., 2018). Therefore, in this experiment, three image patch sizes were selected to evaluate all three algorithms: 72×72 pixels, 128×128 pixels and 256×256 pixels, translating to areas of 0.047 hectares, 0.15 hectares and 0.6 hectares in the real world, respectively. In identifying a particular pixel classification, the whole image patch in which the pixel is centred is fed into the neural network to infer the classification. The identification process outputs binary information about whether a particular pixel is building or not building. This process is then repeated for all required pixels in the satellite image within the study area.

2.1. Experiment design

In this experiment, two independent variables are investigated: 1) the CNN architecture used in the classification and 2) the image patch size. The dependent variable is the accuracy of the classification result, presented through three metrics: 1) Overall Accuracy, 2) Kappa, and 3) Intersection over Union. The hyperparameters for the CNN algorithms were set at predetermined fixed values (as dictated by the hyperparameter tuning process) before the actual training process and were not altered throughout the experiments, so as to ensure that the classification accuracy is only affected by the patch size and the selected CNN algorithm architecture.

2.2. Preparation

For this experiment, the appropriate Keras-TensorFlow framework was installed on the computer, including all the necessary neural network drivers for the Graphics Processing Unit (GPU). In addition, Anaconda environment manager was utilized to organize and simplify Python installation and configuration. Each band in the Worldview-4 imagery needs to be normalized before CNN algorithms process it correctly (Bishop et al., 1995). This normalization can be done by linear minimum–maximum rescaling to ensure that the resulting digital number values range from 0 to 1. This linear rescaling can be accomplished by dividing the raw digital number for the pixels in each channel by 2048 as the dynamic range of Worldview-4 imagery is 11 bits (Digital Globe, 2017), albeit it was delivered in 16-bit-per-channel format. Three different Python scripts were prepared to accomplish these tasks related to the CNN algorithm classification process: 1) building training dataset by automatically generating image patches of the training area based on the manually digitized vectors of the building footprints, 2) training process and model creation, and 3) inference process based on the saved model.

Three sets of image patches of different sizes were generated for 72×72 pixels, 128×128 pixels and 256×256 pixels, with the latter two sizes shown in Figure 4. Each training set with differing image patch sizes was used separately to train the CNN models. In this experiment, a total of 16,000 training image patches for each patch size were generated based on the manually delineated building footprints for all five training areas. The building areas in the five training sites were classified by performing a manual visual interpretation of the satellite image within the training area's boundary. The features interpreted are then digitized into polygonal vectors to signify building and non-building areas. The resulting vectors are then converted into raster data, which will be used to generate sample image patches to be fed into the CNN algorithms during the training process. Each image patch's centre pixel will determine whether that particular patch will be classified as building or non-building. A random sampling procedure was performed to generate sample pixels for the training process for the pixel-based algorithms, RF and KNN.

2.3. Training process

The unweighted models for LeNet-5, VGG-19 and XCeption were prepared in the Python script, instantiated from the built-in CNN models in Keras. These unweighted models will be trained, and the resulting weighted model will be stored in an HDF5 file. The training sessions were performed for each classification algorithm for all the image patch sizes. Since only one computer system is available for research purposes, the training sessions

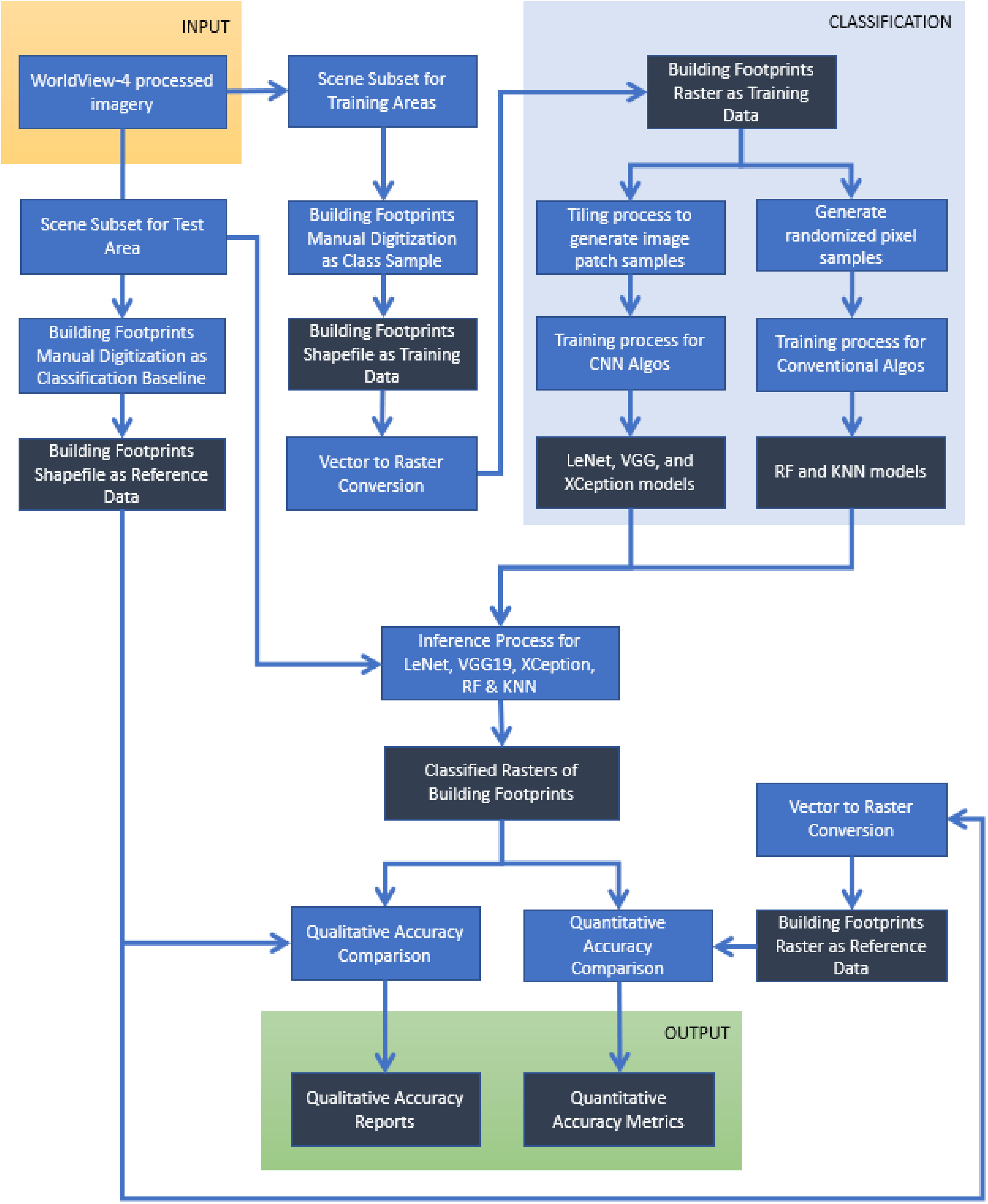
INPUT
WorldView-4 processed imagery
Scene Subset for Training Areas
CLASSIFICATION
Building Footprints Raster as Training Data
Scene Subset for Test Area
Building Footprints Manual Digitization as Class Sample
Tiling process to generate image patch samples
Generate randomized pixel samples
Building Footprints Manual Digitization as Classification Baseline
Building Footprints Shapefile as Training Data
Training process for CNN Algos
Training process for Conventional Algos
Building Footprints Shapefile as Reference Data
Vector to Raster Conversion
LeNet, VGG, and XCeption models
RF and KNN models
Inference Process for LeNet, VGG19, XCeption, RF & KNN
Classified Rasters of Building Footprints
Vector to Raster Conversion
Qualitative Accuracy Comparison
Quantitative Accuracy Comparison
Building Footprints Raster as Reference Data
OUTPUT
Qualitative Accuracy Reports
Quantitative Accuracy Metrics

Fig. 2. Block diagram of the building footprint classification and accuracy evaluation workflow.
Source: own elaboration

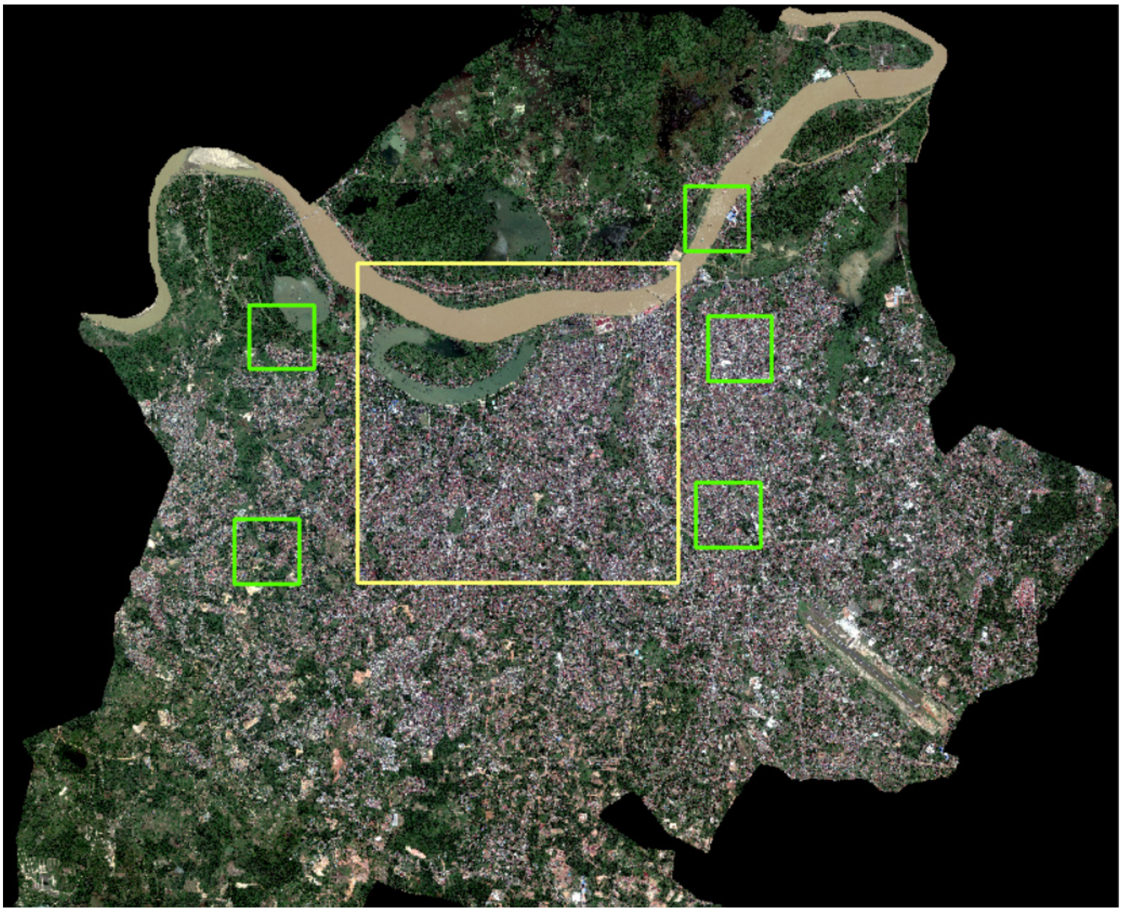

Fig. 3. 2500-hectare test site (yellow rectangle) and training sites (five green rectangles, 100 hectares each)
Source: own elaboration

are repeated for each image patch size set tested. The CNN algorithms' training process's output contains weights for each neural network layer, and these models were stored in HDF5 file format. HDF5 is a high-performance file format for storing heterogeneous data as a model storage format and a container for an organized collection of objects (The HDF Group, 2019) in the Keras-TensorFlow framework. The model varies between architectures, with a larger model size for deeper convolutional neural network architecture.

2.4. Inference process

Based on the model constructed and saved in the form of the HDF5 file produced during the training process, the inference process to produce building classification can then be performed on the entire test area. Since the computation will require a considerable amount of time, the test area was split into 25 tiles at 100 hectares each, with enough overlap between tiles to enable incremental batch processing while minimizing progress loss caused by hardware or power outages. Due to the time constraint and the limitation in computing hardware resources used in this study, not all pixels were classified using CNN algorithms. Instead, only 1 in 16 pixels is processed, thus yielding 1.2-metre resolution in the classification raster. For each pixel-based classification method (RF and KNN), a classification raster with equal resolution (1.2-metre) was generated with no generalization of the classification results.

2.5. Evaluation process

Manual visual identification of the building footprints from the Worldview-4 satellite imagery mosaic was performed in the test area to provide a baseline performance reference. The manual interpretation and classification process was done by a team of professional GIS analysts and took 35 person-hours to complete and an additional 12 person-hours to do quality control. For the classification accuracy

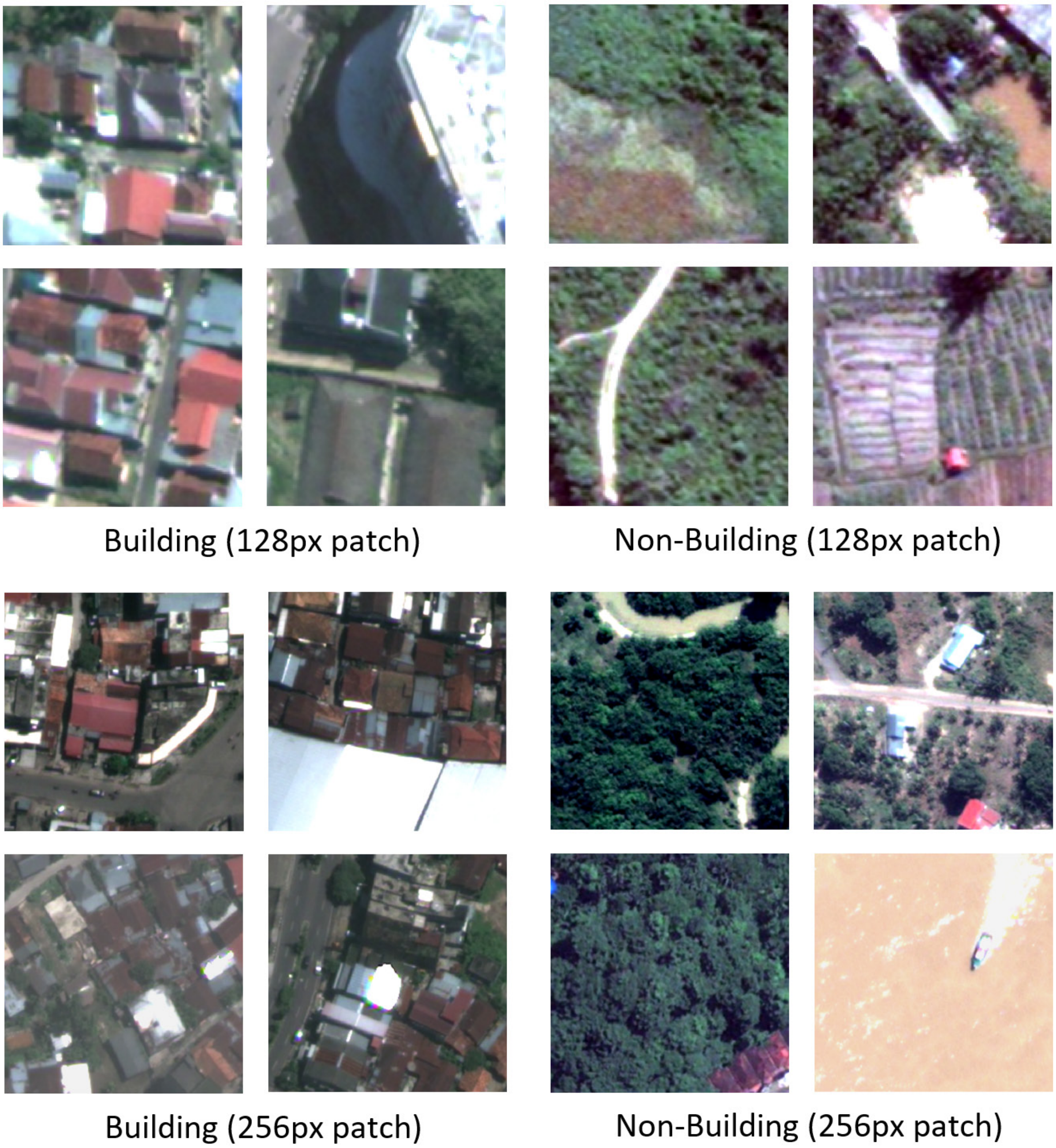

Fig. 4. A few samples of image patches were used in the supervised classification training process. A total of 16,000 image patches similar to these were used
Source: own elaboration

evaluation, a stratified random sampling method was used. A sample of 10,000 test points within the experiment area was selected using a stratified random sampling method. The main characteristic of this sampling method is that it ensures that each classification type within the study area receives proper representation within the sample. The accuracy evaluation will use Overall Accuracy (OA), Kappa coefficient and Intersection over Union (IoU) as the performance metrics, as they are commonly used in Deep Learning literature (Maxwell et al., 2021). The collective accuracy of the classification results can be described using Overall Accuracy, which calculates the proportion of pixels correctly classified as building or non-building. The Kappa coefficient is used as one of the accuracy metrics in this study because of its widespread adoption in the remote-sensing community (Rwanga &

Ndambuki, 2017), even though it has been disputed as a measure of accuracy (Foody, 2020). Intersection over Union (IoU), also known as the Jaccard index, is one of the most popular evaluation metrics for segmentation and object detection tasks and is much more indicative of success for segmentation tasks than pixel-based accuracy metrics (Van Beers et al., 2019). The IoU metric was calculated by comparing the pixels within the inferred building boundaries against the reference data.

3. Research results and discussion

3.1. Quantitative assessment of classification result

All classification results showed overall accuracy (OA) ranging from 84.1% to 88%, Kappa ranging from 0.671 to 0.743, and IoU ranging from 0.638 to 0.716 (Fig. 5). Among the eleven classifiers evaluated in this study, all CNN algorithms performed better than KNN and RF algorithms in all three metrics, with Xception on a 72-pixel image window as the best-performing algorithm. Concerning the image patch size, all classifiers showed varying degrees of OA and Kappa when applied to different image window sizes; however, all CNN classifiers performed better than their conventional counterparts. Table 1 summarizes the accuracy measures (Kappa, OA and IoU) for all evaluated classification algorithms.

3.2. Patch size influence on classification accuracy

In this study, varying patch sizes were tested for all three CNN architectures, and, in all cases, a larger image patch size does not necessarily yield better classification accuracy. In each image patch size, varying performance was observed. XCeption achieved the best accuracy performance at 72-pixel and 256-pixel image window size, while, at 128-pixel window size, VGG performed best.

3.3. Qualitative assessment of classification result

Upon closer inspection of the classification result, the patch-based algorithms tend to over-generalize the building footprints. For example, as shown in Figure 6, the classification boundary did not precisely adhere to the building's outline, as the pixel-based algorithms did. However, areas with no identified buildings are markedly devoid of isolated classification patches or secluded pixels; therefore, minimal generalization is needed if the data is processed for further GIS analysis. In contrast, there is a significant presence of "salt-and-pepper" artefacts in the classification result of the pixel-based classifiers.

3.3.1. High-density urban area

Visual inspection (Fig. 7) showed that the building area classified by the CNN algorithm generally bleeds slightly outwards of the actual building footprint. Small, empty lots of land between tightly packed buildings in the urban area are typically erroneously classified as buildings. The XCeption architecture is capable of differentiating parking lots from buildings, whereas both LeNet-5 and VGG-19 fail to distinguish such cases.

3.3.2. Medium-density urban area

The sub-urban area surrounding the Jambi city centre is characterized by a lower density of buildings compared to the city centre. The classifiers

Table 1. Accuracy metrics comparison, with best values in bold

Accuracy	LeNet-5			VGG-19			XCeption			KNN [4]	RF [5]
Metrics	72px	128px	256px	72px	128px	256px	72px	128px	256px		
OA [1]	0.8746	0.8425	0.8437	0.8637	0.8785	0.8544	**0.8808**	0.8757	0.8716	0.8419	0.8475
KAPPA [2]	0.7234	0.6716	0.6764	0.6841	0.7376	0.6926	**0.7434**	0.7324	0.7177	0.6580	0.6724
IOU [3]	0.6923	0.6617	0.6583	0.6384	0.7083	0.6583	**0.7160**	0.7057	0.6878	0.6384	0.6462

1 Overall Accuracy, 2 Kappa, 3 Intersection over Union, 4 k-Nearest Neighbour, 5 Random Forest
Source: author's own elaboration

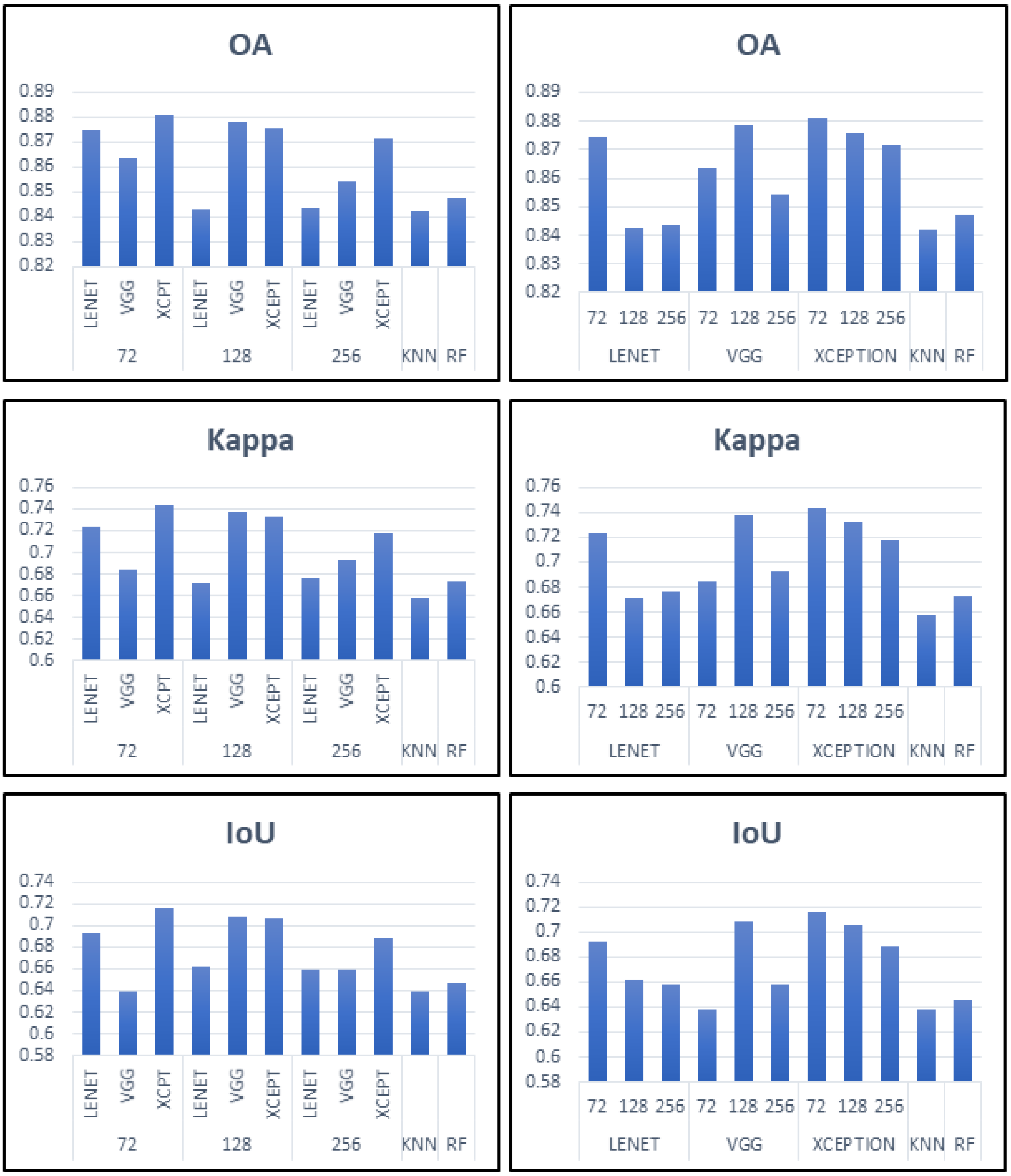

Fig. 5. Accuracy performance chart, grouped by image patch size (left) and by network architecture type (right)
Source: own elaboration

with CNN architecture were able to classify most of the buildings; however, with larger patch size, the classification boundary tends to bleed into the narrow roads (Fig. 8).

3.3.3. Low-density, rural area

Most of the isolated buildings in the rural settlements around Jambi city core were identified correctly by CNN classifiers. However, some buildings with a relatively small footprint and partially occluded by vegetation were misclassified. On the other hand, there are instances where objects that are similar in appearance to a building, such as plastic mulch

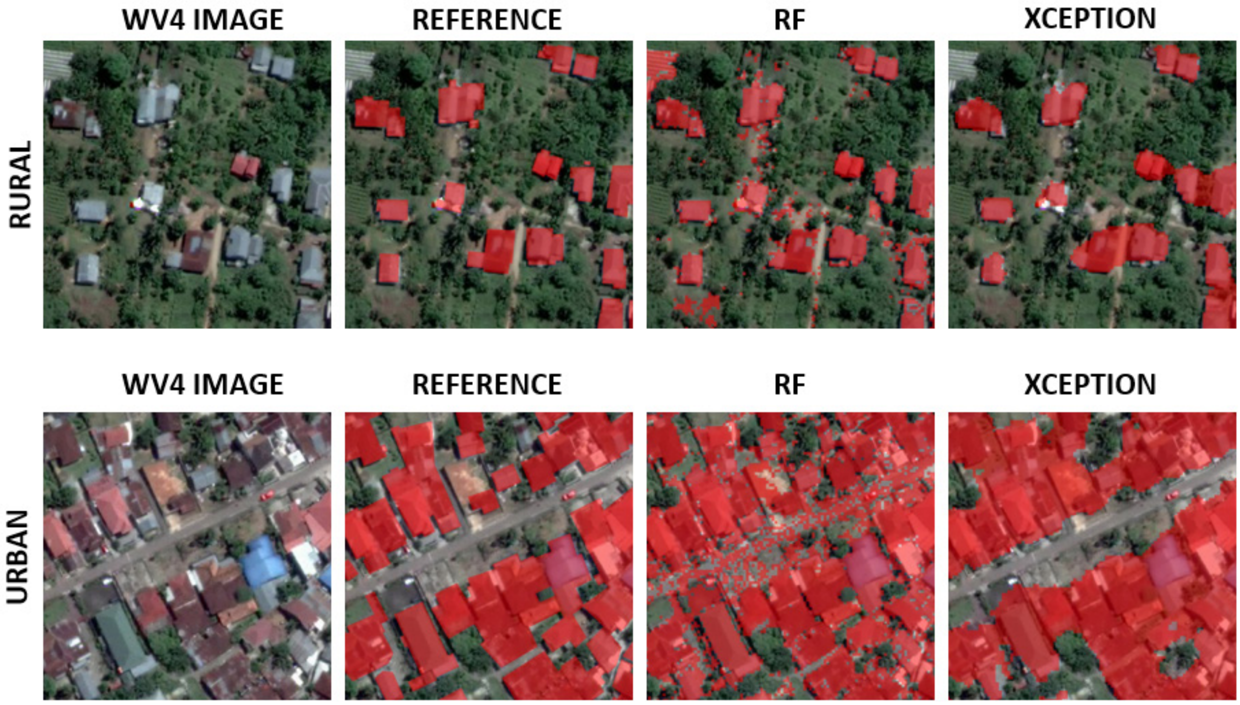

Fig. 6. Classification results in detail, showing both rural and urban settings for select classification algorithms (RF and XCeption).
Source: own elaboration

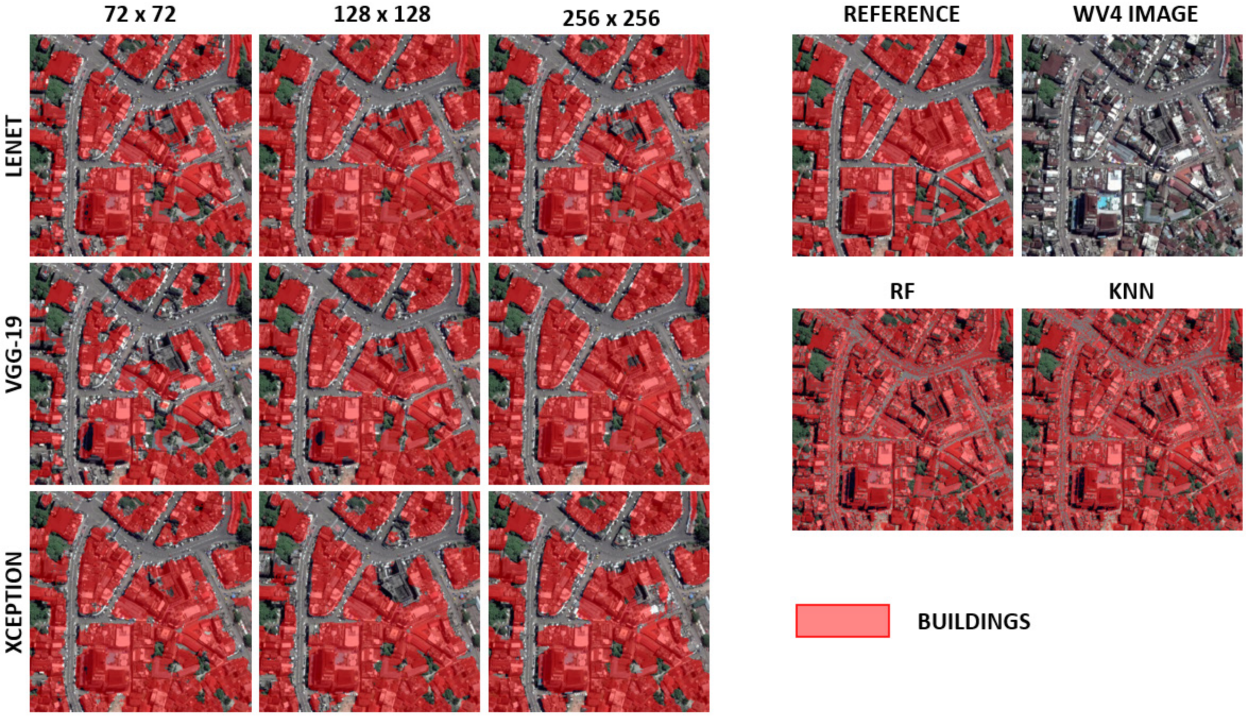

Fig. 7. Classification results on urban areas with high building density
Source: own elaboration

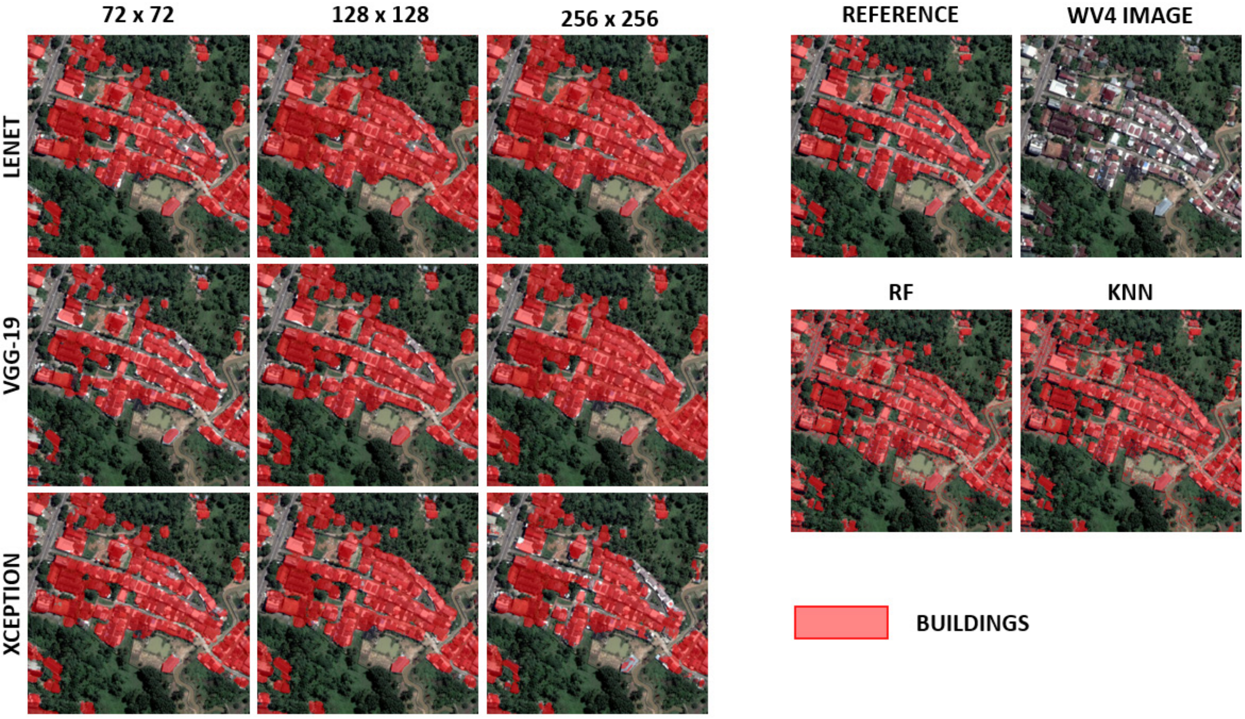

Fig. 8. Classification results on sub-urban areas with low building density
Source: own elaboration

and freshly ploughed fields, were misclassified as buildings (Fig. 9).

The most complex CNN architecture tested in this study – the XCeption architecture – produced the best results compared to other CNN architectures when using 72-pixel image patch size, followed by VGG-19 with 128-pixel image patch size. LeNet architecture yielded the lowest accuracy compared to XCeption and VGG-19; however, as it contains the shallowest convolutional layer depth compared to other CNN algorithms tested in this study, it was much faster to train and to infer, and at 72-pixel image patch size the LeNet architecture still produced better results compared to conventional, pixel-based classification algorithms. There is a trade-off between the hardware requirements, data processing timeframe limitations and the expected accuracy levels.

This experiment has shown that the Deep Learning classification algorithm is a viable alternative for generating accurate building footprints maps, in line with previous studies (Längkvist et al., 2016; Hamwood et al., 2018; Kattenborn et al., 2019;). Furthermore, this experiment has shown that the CNN classifier produces better accuracy than conventional pixel-based classification algorithms, which agrees with previous studies (Kussul et al., 2017; Zhang et al., 2018; Pan et al., 2020). However, even though the quantitative results are satisfactory, the CNN algorithms fail to infer some specific areas correctly; for example, swimming pools, large bridges and stacks of shipping containers were classified as buildings. These objects were not featured in the training area; therefore, all of the CNN algorithms misclassified them. The proper training dataset remains a critical aspect determining the classification accuracy (Kavzoglu, 2009; Millard & Richardson, 2015); this study shows that even the highest-performing CNN architecture fails to correctly identify objects not featured in the training area.

Paved surfaces such as roads and parking lots, even smaller ones, were successfully identified and classified as non-building by the CNN algorithms. In contrast, in pixel-based classification algorithms, almost any paved surface with a similar radiometric spectral signature as buildings in the training area were classified as buildings. Spatial and contextual data can provide valuable information about the shape of different structures, and such information reduces the classification uncertainty that arises when only spectral information is taken into account and also helps to address the "salt and pepper" artefacts of the resulting classification map (Tarabalka et al., 2018). Nevertheless, as the spatial and contextual scope presented to the pixel-based algorithms was only limited to one single pixel with 30-cm resolution, as in the case of WorldView-4

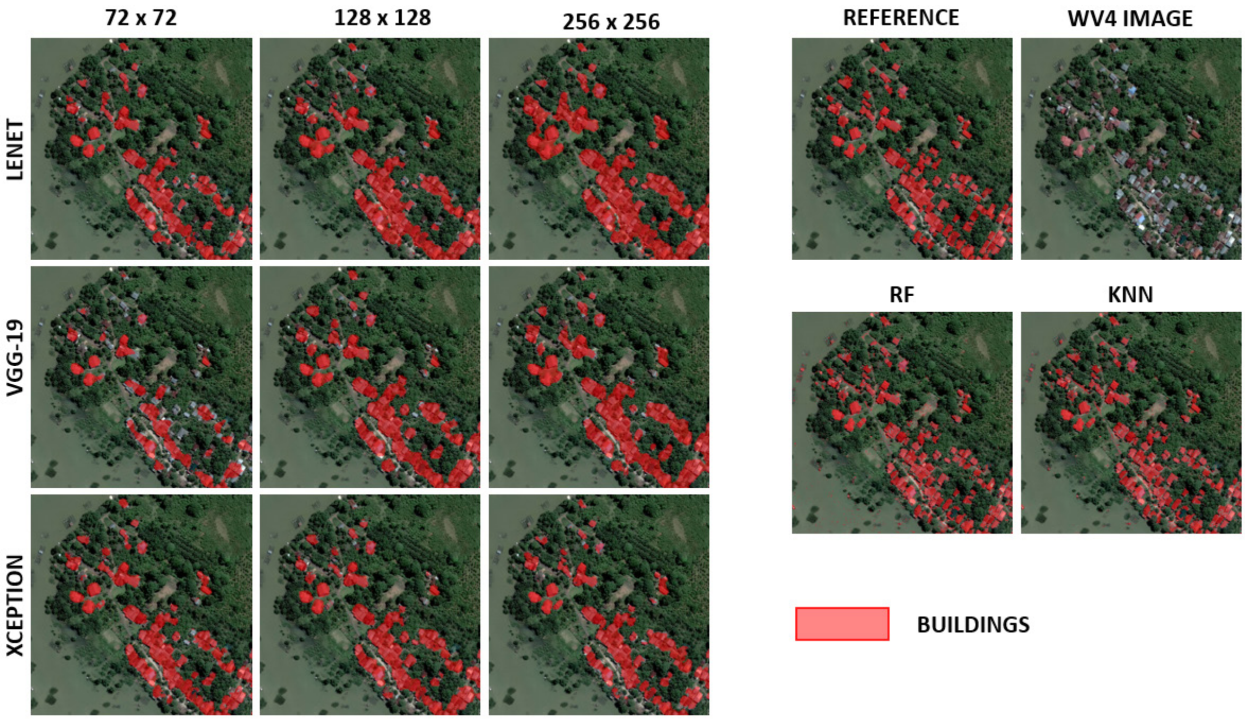

Fig. 9. Classification results on sub-urban areas with low building density
Source: own elaboration

imagery, much noise was generated in the form of spurious classified pixels or "salt and pepper" appearance in the classification result of the RF and KNN algorithms. This noisy result necessitates a cartographic generalization process before the results can be used further. On the other hand, these spurious small bits of incorrect classification were not observed in the output of the three CNN classifiers evaluated in this study, since the CNN classifiers were fed with more spatial and contextual data within the image patches. However, this study showed that larger image patch size does not always result in higher classification accuracy, as each CNN architecture showed diminishing returns with increasing image patch size.

4. Conclusions

In building classification from high-resolution satellite imagery, the CNN algorithms have significantly and consistently outperformed conventional classification methods, such as Random Forest and k-Nearest Neighbour, in terms of classification accuracy, especially in the study area. However, this higher accuracy comes with the higher cost of a more advanced hardware set-up, which will cost more to procure and take much longer to produce the classification raster.

More sophisticated CNN architectures with deeper layers can differentiate more subtle patterns and variations of surficial features captured in the satellite imagery, which yields better classification accuracy than pixel-based algorithms. Pixel-based classification algorithm depends on band combination to perform classification for each pixel. Therefore, when given only three natural colour bands, pixel-based algorithms may not produce accurate results, specifically when semantic information can be inferred from the spatial pattern of the surrounding pixels. On the other hand, the CNN algorithms can capture such information because they use image patches instead of single pixels for the classification process. However, larger image patches do not necessarily increase accuracy, as this study has proven. Different CNN architectures perform best at different image patch sizes.

Whereas the CNN algorithms require more time to train and classify, most of the time spent on the process is mostly computing time with minimum human interaction; this is in stark contrast with manual classification, which requires full-time human work. However, although lower in terms of accuracy, conventional pixel-based classification algorithms can perform much faster than their patch-based counterparts while requiring lower hardware specifications to run. Further research can be done for other unique areas in different geographic regions by utilizing other satellite imagery sensors, ranging from medium-resolution sensors such as Landsat-8 and Sentinel-2 to high-resolution sensors, such as Pleiades and WorldView satellite series.

Acknowledgements

The Jambi City Worldview-4 satellite image dataset was provided by courtesy of the Head of Spatial Planning, Department of Public Works and Spatial Planning, Jambi Municipality Government, Indonesia. The authors are grateful to the Department of Geography, Faculty of Mathematics and Natural Science, Universitas Indonesia, for facilitating the research for this study.

References

Aroma, J. & Raimond, K. (2016). An Overview of Technological Revolution in Satellite Image Analysis. *Journal of Engineering Science and Technology Review*, 9(4): 1–6.

Ayala, C., Sesma, R., Aranda, C. & Galar, M. (2021). A deep learning approach to an enhanced building footprint and road detection in high-resolution satellite imagery. *Remote Sensing*, 13(16): 1–21.

Bishop, C.M., Bishop, P.N.C.C.M., Hinton, G. & Press, O.U. (1995). *Neural Networks for Pattern Recognition*. UK: Clarendon Press.

Chawda, C., Aghav, J. & Udar, S. (2018). Extracting Building Footprints from Satellite Images using Convolutional Neural Networks. *2018 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2018*, September 2018, 572–577.

Chollet, F. (2014). Xception: Deep Learning with Depthwise Separable Convolutions. *Computer Vision and Pattern Recognition*, 1251–1258.

Chollet, F. (2015). Keras. Available at: https://keras.io/ (Accessed 10 April 2022).

Digital Globe. (2017). WorldView-4 Data Sheet. Available at: https://dg-cms-uploads-production.s3.amazonaws.com/uploads/document/file/196/DG2017_WorldView-4_DS.pdf (Accessed 10 April 2022).

Fatima, S.A., Kumar, A. & Raoof, S.S. (2021). Real Time Emotion Detection of Humans Using Mini-Xception Algorithm. *IOP Conference Series: Materials Science and Engineering*, 1042(1): 012027.

Fitri, S.H. & Sumunar, D.R.S. (2019). The Direction of Development of Jambi City Based on Flood Disaster Mitigation. *IOP Conference Series: Earth and Environmental Science*, 271(1).

Gikunda, P.K. & Jouandeau, N. (2019). State-of-the-Art Convolutional Neural Networks for Smart Farms: A Review. *Advances in Intelligent Systems and Computing*, 997: 763–775.

Government of Indonesia (2013). Peraturan Pemerintah Republik Indonesia Nomor 8 Tahun 2013 Tentang Ketelitian Peta Rencana Tata Ruang (Government of Republic Indonesia Regulation Concerning the Map Accuracy for Spatial Planning – in Indonesian), PP No 8 Tahun 2013, 1.

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

Hardiani, H. & Lubis, T.A. (2017). Analysis of leading sector of Jambi City. *Jurnal Perspektif Pembiayaan Dan Pembangunan Daerah*, 5(1): 1–12.

Hu, Y., Zhang, Q., Zhang, Y. & Yan, H. (2018). A deep convolution neural network method for land cover mapping: A case study of Qinhuangdao, China. *Remote Sensing*, 10(12): 1–17.

Kattenborn, T., Eichel, J. & Fassnacht, F.E. (2019). Convolutional Neural Networks enable efficient, accurate and fine-grained segmentation of plant species and communities from high-resolution UAV imagery. *Scientific Reports*, 9(1): 1–9.

Kavzoglu, T. (2009). Increasing the accuracy of neural network classification using refined training data. *Environmental Modelling and Software*, 24(7): 850–858.

Khan, A., Sohail, A., Zahoora, U. & Qureshi, A.S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53(8): 5455–5516.

Kussul, N., Lavreniuk, M., Skakun, S. & Shelestov, A. (2017). Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. *IEEE Geoscience and Remote Sensing Letters*, 14(5): 778–782.

Längkvist, M., Kiselev, A., Alirezaie, M. & Loutfi, A. (2016). Classification and segmentation of satellite orthoimagery using convolutional neural networks. *Remote Sensing*, 8(4): 329.

Lecun, Y., Bottou, L., Bengio, Y. & Ha, P. (1998). Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, November, 1–46.

Li, L., Qiang, Y., Zheng, Z. & Zhang, J. (2019). Research on the Relationship between the Spatial Resolution and the Map Scale in the Satellite Remote Sensing Cartographies.

International Conference on Modeling, Analysis, Simulation Technologies and Applications (MASTA 2019), 194–199.

Luo, L., Li, P. & Yan, X. (2021). Deep learning-based building extraction from remote sensing images: A comprehensive review. *Energies*, 14(23): 1–25.

Maxar. (2020). 2019 Annual Report. In 2019 Annual Report (Issue January: 26). Available at: https://s22.q4cdn.com/683266634/files/doc_financials/2019/ar/Maxar-2019-AR-Web-PDF.pdf (Accessed 10 April 2022).

Maxwell, A.E., Warner, T.A. & Guillén, L.A. (2021). Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—Part 1: Literature Review. *Remote Sensing*, 13(13): 2450.

Millard, K. & Richardson, M. (2015). On the importance of training data sample selection in Random Forest image classification: A case study in peatland ecosystem mapping. *Remote Sensing*, 7(7): 8489–8515.

Muhammad, U., Wang, W., Chattha, S.P. & Ali, S. (2018). Pre-trained VGGNet Architecture for Remote-Sensing Image Scene Classification. *Proceedings - International Conference on Pattern Recognition*, 2018-Augus(August), 1622–1627.

Pacheco, A.D.P., Junior, J.A.D.S., Ruiz-Armenteros, A.M. & Henriques, R.F.F. (2021). Assessment of k-nearest neighbor and random forest classifiers for mapping forest fire areas in central portugal using landsat-8, sentinel-2, and terra imagery. *Remote Sensing*, 13(7): 1–25.

Pal, M., Akshay, Rohilla, H. & Teja, B.C. (2020). Patch Based Classification of Remote Sensing Data: A Comparison of 2D-CNN, SVM and NN Classifiers. CoRR, abs/2006.1. Available at: https://arxiv.org/abs/2006.11767 (Accessed at: 10 April 2022).

Pan, Z., Xu, J., Guo, Y., Hu, Y. & Wang, G. (2020). Deep learning segmentation and classification for urban village using a worldview satellite image based on U-net. *Remote Sensing*, 12(10): 1–18.

Rwanga, S.S. & Ndambuki, J.M. (2017). Accuracy Assessment of Land Use/Land Cover Classification Using Remote Sensing and GIS. *International Journal of Geosciences*, 08(04): 611–622.

Shah, T.N., Khan, M.Z., Ali, M., Khan, B. & Muhammad, H. (2018). Critical Analysis of Six Frequently Used Classification Algorithms. *University of Swabi Journal*, 2(2): 36–40.

Simonyan, K. & Zisserman, A. (2018). Very Deep Convolutional Networks For Large-Scale Image Recognition. *American Journal of Health-System Pharmacy*, 75(6): 398–406.

Sultana, F., Sufian, A. & Dutta, P. (2018). Advancements in image classification using convolutional neural network. *Proceedings - 2018 4th IEEE International Conference on Research in Computational Intelligence and Communication Networks*, ICRCICN 2018, 122–129.

Tarabalka, Y., Moser, G., Giorgi, A.D.E., Fang, L., Chen, Y., Chi, M., Serpico, S.B. & Benediktsson, J.Ó.N.A. (2018). New Frontiers in Spectral- Spatial Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, September 2018.

Thanh Noi, P. & Kappas, M. (2017). Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. Sensors (Basel, Switzerland), 18(1).

The HDF Group. (2019). HDF5 User's Guide. Available at: https://support.hdfgroup.org/.

Tobler, W. (1987). Measuring Spatial Resolution. *Beijing Conference on Land Use and Remote Sensing*, July, 12–16. Available at: https://www.researchgate.net/publication/.

Van Beers, F., Lindström, A., Okafor, E. & Wiering, M.A. (2019). Deep neural networks with intersection over union loss for binary image segmentation. *ICPRAM 2019 - Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*, 438–445.

Xiao, J., Wang, J., Cao, S. & Li, B. (2020). Application of a Novel and Improved VGG-19 Network in the Detection of Workers Wearing Masks. *Journal of Physics: Conference Series*, 1518(1).

Zhang, M., Lin, H., Wang, G., Sun, H. & Fu, J. (2018). Mapping paddy rice using a Convolutional Neural Network (CNN) with Landsat 8 datasets in the Dongting Lake Area, China. *Remote Sensing*, 10(11): 1840.

HDF5/doc/UG/HDF5_Users_Guide-Responsive%20HTML5/index.html (Accessed 10 April 2022).

291877360_Measuring_spatial_resolution (Accessed 10 April 2022)