Urban Land Cover Classification Using Deep Neural Networks Based on VHR MS Image and DSM

Mohamed Fawzy[1,2,*], Attila Juhasz[1], Arpad Barsi[1]

[1] Department of Photogrammetry and Geoinformatics, Faculty of Civil Engineering, Budapest University of Technology and Economics, Műegyetem rkp. 3, H-1111 Budapest, Hungary {mohamed.fawzy, juhasz.attila, barsi.arpad}@emk.bme.hu

[2] Civil Engineering Department, Faculty of Engineering, South Valley University, 83523 Qena, Egypt, mohamedfawzy@eng.svu.edu.eg

* Corresponding author

Abstract: Urban environment presents one of the most challenging research fields for remote sensing data analysis tasks due to the wide range of land cover materials, and the variety of land use classes. Urban feature extraction, e.g. buildings and roads, has a significant impact on several applications such as urban planning and management, monitoring human activities, and change detection. Effective feature extraction procedures require both high-quality data and reliable processing methodology. Recent progress in remote sensing techniques offers a wide range of Multi-Spectral (MS) images and Digital Surface Models (DSMs) for urban studies. Deep Learning (DL) approaches, especially Convolutional Neural networks (CNNs), have a notable performance in handling large amounts of data with the advantage of mapping the relationship between high-dimensional and nonlinear features. The current research employs the U-NET to develop a CNN model for urban feature extraction from multi-spectral images and DSMs. The proposed U-NET is trained, validated, tested, and applied for image semantic segmentation using integrated WorldView-2 multi-spectral image and LiDAR DSM. The classified urban features are subsequently refined based on the elevation values of the DSM. Four accuracy indices: correctness, completeness, quality, and overall accuracy are calculated to evaluate the obtained outcomes and check the model stability. Building extraction has attained an overall accuracy of 69.1% and 89.9% for classified and refined images respectively, whereas road extraction has obtained an overall accuracy of 89.9% and 90.7% for classified and refined images respectively. The U-NET model has achieved promising outcomes for image semantic segmentation, while the DSM added a notable enhancement during refinement.

Keywords: Convolutional neural networks; VHR multi-spectral images; DSMs; Urban feature extraction

1 Introduction

Cities are under increasing pressure due to the economic, social, and environmental spheres since the extraordinary rise in global urbanization, which is predicted to reach roughly 70% by 2050. Rapid urbanization has exacerbated several issues that threaten the sustainability of metropolitan areas. Sustainable development is the state in which ecological regeneration and sustainable economic growth coexist in harmony aiming to achieve future municipal goals without sacrificing social cohesiveness, environment, or human well-being [1]. Urban agglomerations face long-term challenges because of future demographic shifts, climate changes, and structural major developments. Natural disasters, pollution, inadequate infrastructure, and unequal living conditions pose severe threats to the sustainability of the environment, infrastructure, and urban societies. Urban studies are continuously looking for novel data and effective methodologies to meet these difficulties. Remote sensing strategies support data-driven decision making and efficient management in urban planning and policymaking [2]. Land cover analysis is an essential task for creating Land Use Land Cover (LULC) maps for urban applications using remotely sensed Very High Resolution (VHR) data. Using VHR images has several advantages including timely information, low costs, and the synoptic view of urban land cover. Furthermore, Digital Surface Models (DSMs) are invaluable sources for elevation information about urban features including building, road, and vegetation classes. Combining DSM and VHR satellite imagery is recommended to improve the land cover classification [3]. DSM and hyperspectral dataset integration improves the overall accuracy of the combined datasets better than the individual ones. The added features, such as elevation data from DSMs, offer a more comprehensive representation for the creation of precise maps and improve the classification accuracy of particular urban applications [4]. To keep up with the exponential growth in collected data, it is essential to develop automatic methods for assessing VHR images and DSMs where Artificial Intelligence (AI)-based strategies can be useful. AI systems are capable of data analysis, pattern recognition, and well-informed prediction. Semantic segmentation networks based on Deep Learning (DL), especially Convolutional Neural Networks (CNNs), showed encouraging results for automatic image classification at pixel level. CNNs have achieved cutting-edge achievements in image classification thanks to their capacity to extract high-level spatial data using integrated connections. However, obtaining constant accuracy requires a large enough amount of labeled data, which is considered the real challenge of the DL-based methods [5]. Recently, deep learning techniques have gained considerable popularity in the fields of computer vision, remote sensing, and pattern recognition. Ma, L., et al. [6] covered almost all applications of using artificial intelligence in remote sensing, from preprocessing to mapping, and presented the key DL ideas relevant to remote sensing using 200 published works over the previous years. Neupane, B. et al. [7] concentrated on urban remote sensing imagery and reviewed articles addressing research questions, data sources,

data pre-processing techniques, and the details of architectural training. The outcomes fixed several problems and showed that deep learning outperforms traditional methods in terms of accuracy. CNNs represent a deep learning system that learns representations using training samples in a hierarchical way, controls and shares the weight related to the training data, generalizes, optimizes the parameters, and minimizes the errors with a better level of automatic feature extraction. The network uses many convolutional layers for learning how to detect image features and improves the labeling process over time. CNNs offer an appropriate architecture for analyzing extremely high-resolution datasets and empowers the capacity to automatically learn contextual characteristics from input images [3]. Generally, neural networks are composed of three different types of layers: input, hidden, and output layers. The hidden layers are usually composed of a convolutional, pooling, fully connected, and normalizing layer sequence. Each algorithm is unique according to the architecture it employs and the purpose for which it is designed [8]. CNNs such as U-Net [9], GoogLe-Net [10], Alex-Net [11], Dense-Net [12], VGG-Net [13], and Res-Net [14] have demonstrated strong performance in multimedia image analysis tasks [15].

2 Research Problem and Objectives

Land use identification in urban environments is particularly challenging due to the significant diversity of various land cover classes and materials which presents various obstacles for remote sensing applications and data interpretation [16]. Moreover, the Multi-Spectral (MS) images experience notable similarities between the spectrum reflectance of different classes, due to the wide range varieties of used materials [17]. The state-of-the-art presents a set of traditional methods for land cover classification using multi-spectral VHR satellite data like pixel-based image analysis and object-based image analysis. However, traditional approaches require a long time for processing, consume more computer power and memory, and involve extensive human efforts to finish classification tasks. The accuracy and effectiveness of standard techniques are insufficient to meet the requirements of practical applications, like land management and sustainable urban planning. With the increasing availability of remotely sensed data, gathering, organizing, and analysing urban information using conventional manual methods and standard software has become a complicated task. Deep learning techniques are well known for their emphasis on maximizing the extracted urban land cover features learning from large datasets of satellite images [18]. The main aim of the presented work is to investigate CNNs for urban feature extraction using VHR satellite imagery and DSMs focusing on six main classes: water, grasslands, trees, bare land, roads, and buildings. Classification results are refined using the elevation values of the DSM to determine whether the fused DSM and MS image produces noteworthy results for road and building extraction.

3 Methodology

The recommended approach entails developing a CNN model to handle several variables based on VHR multi-spectral satellite image and DSM for land cover classification in urban environments. The study objective is achieved through the following methodology where six steps have been performed and their results are evaluated (Figure 1).

- A suitable study area is selected with available VHR satellite image and DSM.
- Image preprocessing is applied to rectify and unify the coordinates of the VHR image and the DSM.
- Data preparation is performed to fuse the multi-spectral image with the DSM as an input image for the neural network.
- A CNN model is developed, trained, and tested for image classification into six main classes.
- Extracted buildings and roads are refined to remove the confusion with other classes and enhance the outcome's accuracy.
- The extracted features are assessed considering four indices: completeness, correctness, quality, and overall accuracy.

4 Experimental Works

4.1. Study Area and Data Used

Remote sensing data is highly advantageous for urban studies due to its availability, global coverage, wide range of scales and resolutions, and ease of application [2]. Spectral and spatial information from remote sensing imagery is plentiful to gain a comprehensive understanding of urban features. Multi-spectral satellite images are employed in urban applications, e.g. LULC classification frequently using the Red-Green-Blue (RGB) and Near-Infrared (NIR) bands [19]. Water class can be identified using green, blue, and NIR bands. Vegetation classes are detected employing bands in the red and NIR spectrums. Bare soil class is extracted by the red, green, yellow, and NIR bands. Urban features, buildings and roads, could be classified using red, yellow, blue, coastal blue, red, and NIR bands [17].

4.1.1. Multi-Spectral Satellite Image

High-resolution images captured by the WorldView-2 satellite are extensively available and commonly utilized. The European Space Agency, in partnership with European Space Imaging, has made the WorldView-2 dataset accessible, encompassing the most densely populated areas of Europe [20]. A WorldView-2 satellite image of Heiligensee near Berlin, (Figure 2-a) is used containing a MS channel with eight bands of spatial resolution 2.00 m: three red, green, blue, two NIR, coastal blue, yellow and red-edge covering spectral range of 400 nm - 1200nm. In addition, the panchromatic channel has a spectral range of 450–800 nm, and spatial resolution of 0.46 m GSD at nadir (0.52 m at 20° off-nadir). All channels have a geolocation accuracy of less than 3 meters in the absence of any Ground Control Points (GCPs). The WorldView-2 product is orthorectified according to the Standard(2A)/View Ready Standard (OR2A). This level of preprocessing includes orthorectification, projection and resampling, and photogrammetric analysis required for feature extraction in geographic information systems. The WGS84 reference frame serves as the foundation for the geodetic coordinate system, which projects coordinates expressed from latitude and longitude to UTM, Zone 33 coordinates (EPSG Code: 32633) [21].

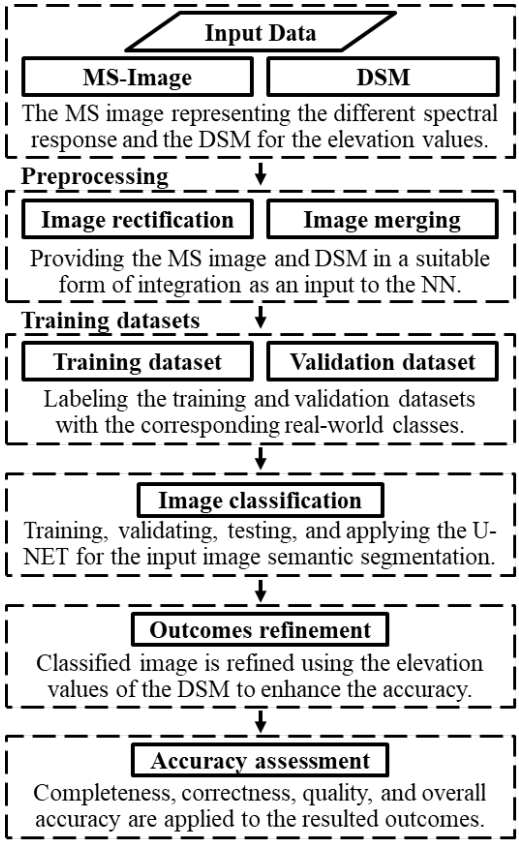


Figure 1
Flow chart of the proposed methodology

4.1.2. DSM Data Source

Digital Elevation Models (DEMs) constitute one of the most important digital spatial datasets for several urban applications. Two terms are commonly used in connection with DEMs: Digital Surface Models (DSMs) and Digital Terrain Models (DTMs). A DTM is a more broader term refers to a DEM with one or more types of topographical information, including morphological components, drainage patterns, and soil properties [22]. The DEM deals with a single type of topographical information, such as height representing the land surface with no trees, buildings, or other non-ground items present. Meanwhile, a DSM shows the tops of all objects, including trees, buildings, and the bare land. Several free global digital elevation model sources, like the Shuttle Radar Topography Mission (SRTM) and the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), are accessible with spatial resolutions of typically around

30 m [23]. However, their application in classifying metropolitan areas is meaningless due to the course resolution. A DSM can be created, for example, by interpolating mass points including the elevation aspects of natural or man-made elements, such as trees, buildings and roads. Light Detection and Ranging (LiDAR) is the primary technique used to supply data for DSMs. LiDAR technology provides quick and high-resolution surface elevation data acquisition for a wide range of urban applications [24]. As the resolution of the available DSM did not align with our initial expectations, the DSM was developed using LiDAR data. The original airborne laser-scanned data was obtained from the geoportal webpage of Berlin [25], which offers a comprehensive array of services including spatial datasets, attribute data, geodata catalog, metadata, and more. A DSM with a spatial resolution of 1.00 m (Figure 2-b) is used and integrated with the MS image to enhance the urban feature extraction. The resolution of the derived DSM is like the resolution of the resampled satellite image. A concise overview of the LiDAR data is provided in Table 1.

(a) MS image (RGB view) (b) DSM

Figure 2
WorldView-2 multi-spectral image and DSM of the selected study area

Table 1
The characteristics of the sample LiDAR dataset

Format	LAS 1.4
Point density	9.8 points / m^2
Section	1 × 1 km
Flight date	24. and 25. 02.2021, 02.03.2021
Scale of survey	1:1000
Coordinate system	ETRS89 / UTM zone 33N

4.1.3. Study Area

Urban study areas are composed of similar and complex features such as buildings, roads, water, trees, grasslands, and bare soil. To fulfill the research objectives, it was imperative to select a study area with access to the necessary datasets. Given that the primary aim was to identify buildings and roads, the data resolution needed to approximate 1.00 m. A significant portion of freely downloadable satellite images meet these requirements. Conversely, the available free digital surface models (DSM) typically offer lower resolutions, ranging from 20 to 30 m. Consequently, an alternative approach was necessary to obtain a suitable digital surface model. Fortunately, owing to evolving data policies, an increasing number of freely available airborne laser-scanned data with resolutions of 1.00 m or higher is accessible in several countries. Considering the available options, the chosen study area is situated in Heiligensee, an agglomeration settlement within the German capital, Berlin (Figure 3). The area covers 0.43 km^2. This region exemplifies a typical agglomeration zone characterized by detached housing, paved roads, water, vegetation and bare lands.

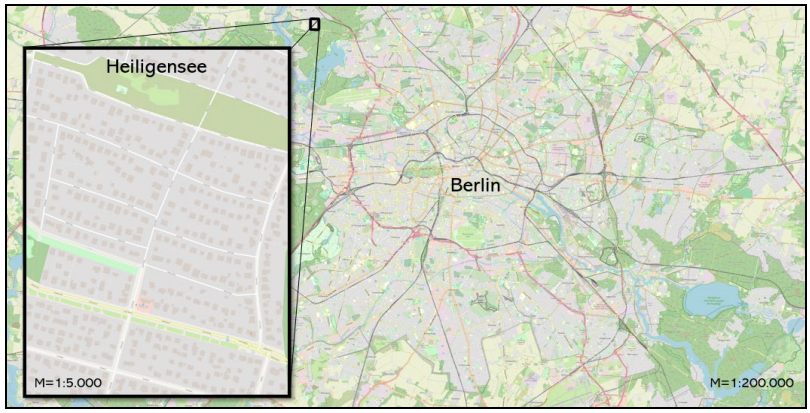


Figure 3
A window of the selected study area map

4.2. Data Preprocessing

Preprocessing of the remote sensing data is required to provide the used images and DSM in a more suitable form considering the neural network input. The preprocessing is applied to the collected data through two steps: geometric registration and data fusion. For the input data registration, the satellite image is registered to a known georeferenced frame to compensate the variations in the geometric information of the image to match the real-world. The DSM is transformed into a grid and handled similarly as an image where image to image registration is required. Moreover, the MS image and the DSM are obtained from

different sources, in which a variety of factors including the perspective of the sensor optics, the motion of the scanning system, the platform altitude, velocity, and Earth rotation result in geometric errors in the final deliverables. Therefore, image-to-image rectification is applied through several transformation processes to minimize the geometric distortion and register the MS image and the DSM to the same coordinate system. The third-order polynomial transformer is used for the registration process using Erdas Imagine software [26]. The multi-spectral image contains a PAN band in high spatial resolution in addition to multi-spectral bands in coarse spatial resolution. To create a fine spatial image with different spectral data, the high spatial resolution from the PAN band is injected into the multi-spectral image using a pan-sharpening procedure. The Principal Component Analysis (PCA) technique is applied, using Erdas Imagine software, for the pan-sharpening process due to its advantages of preserving the number of the eight multi-spectral bands the same before and after the fusion [27].

4.3. Image Classification using Deep Neural Networks

4.3.1. Classification Model

The image classification process involves a series of consecutive phases, starting with reading the input data and ending with obtaining the classified outcomes. The proposed model (Figure 4) considers reading the collected MS image and DSM. A geometric registration is applied to rectify the DSM coordinating system according to the MS spatial information. The corrected inputs are merged into a nine-band input image (8 multi-spectral bands of the MS image and one layer for the elevation data from the DSM). The merged image is used in collecting labeling data for the neural network. The labeling datasets are used to train, validate, and test a U-Net for image semantic segmentation into the output urban classes. The U-Net convolutional neural network adopts a training strategy that mainly relies on data augmentation to maximize the usage of the available labeled samples for efficient and precise image segmentation [9]. The model outcome is a classified image with the main urban classes: water, grasslands, trees, bare land, buildings and roads.

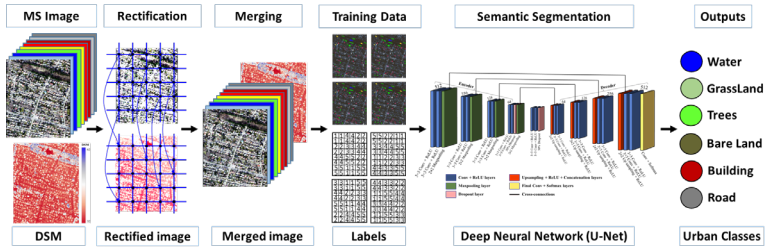


Figure 4
The proposed classification model using convolutional neural network

4.3.2. U-NET Model Architecture

Convolutional neural networks are presented by a series of different layers: convolution layers, nonlinear activation function layers, like Rectified Linear Unit (ReLU), pooling layers, dropout layers, and softmax layers. Convolutional layers use a network of learnable convolutional filters (kernels) to identify high-level features from training datasets. The nonlinear activation function layers are added to exclude the negative values of the activation results by the first operation. The pooling layers consider the maximum value in a local patch for the output to combine the correlated and statistically similar features into one object which minimizes the spatial size of the output and controls the overfitting. The dropout layers randomly eliminate (to zero) a portion of the input units, usually between 20% and 50%, of the previous layer to prevent overfitting. The softmax layers are used to predict the unique classes and assign the most likely label to each pixel in the image by calculating the difference between validation data and training datasets [3]. The suggested model employs a U-NET with the standard architecture consisting of an encoder section, mid-layer, and decoder section (Figure 5) [28].

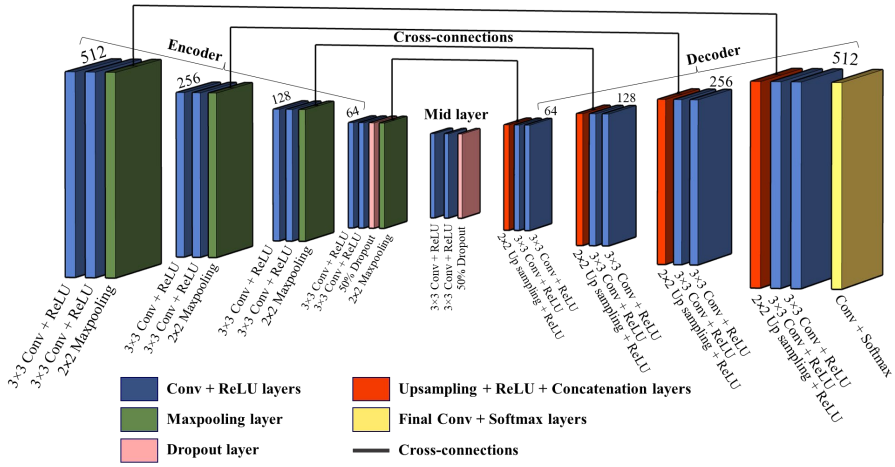


Figure 5
The applied U-Net architecture

The encoder section uses 4 blocks, each block has double 3×3 convolutional + ReLU layers and a 2×2 maxpooling layer. While a dropout layer with a 50% probability value is added before the maxpooling layer of the last block. The mid-layer presents double 3×3 convolutional + ReLU layers and finally a 50% dropout layer. The decoder section is similar to the encoder with 4 blocks each one containing a 2×2 upsampling convolutional + ReLU layers and double 3×3 convolutional + ReLU layers. Following the last block, a convolution layer and six 1×1 softmax layers are connected to present the classification outcomes.

Finally, cross-connections are used to link each block in the encoder to the related one in the decoder. The neural network considers the properties of the incoming 9-band multi-spectral image and DSM data to be classified into 6 classes. Initially, the weights and biases for the convolutional layers were determined using a random number generator that produces uniformly distributed random numbers.

4.3.3. Data Labeling and Model Training

MATLAB Image Labeler application is used to pick training data with the interactive ground truth classes in the input images. Two data sets are created for the training and validation phases with sufficient and efficient samples and their corresponding labels. After multiple trials, a total of 16,000 (128×128 pixels) training patches were chosen since a smaller number resulted in overfitting and poor accuracy, while a larger number required a large investment of time, computing memory and processing resources. 2000 validation patches were discovered given that the standard ratio used to range between 10% and 15% of the dataset. As such, there appears to be a serious challenge in effectively implementing CNN model training on devices with limited resources. For the implemented U-NET, the cross-entropy loss function is determined by comparing the predictions with the related real labels to measure the performance of a classification model. Throughout the training process, all parameter and weight adjustments are updated layer by layer. The Stochastic Gradient Descent with Momentum (SGDM) optimizer is used for the training as an effective optimization method. SGDM aids in the proper acceleration of gradient vectors, resulting in a faster convergence. Stochastic gradient descent is then used in batch normalization to enhance training and minimize loss prior learning each convolutional layer [8]. The used momentum and learning rate are 0.9 and 0.05, respectively. The training process includes a validation phase. Training focuses on adjusting the model parameters, while validation offers an objective assessment of the model. The entire calculations are arranged into sixteen-piece minibatches. A minibatch is a training set subset to update the weights and assess the gradient of the loss function. Training and validation progress and loss are monitored graphically to track the segmentation network parameter change process (Figure 6). The model has an obvious training, validation accuracy and loss progression with a final validation of 99.00%.

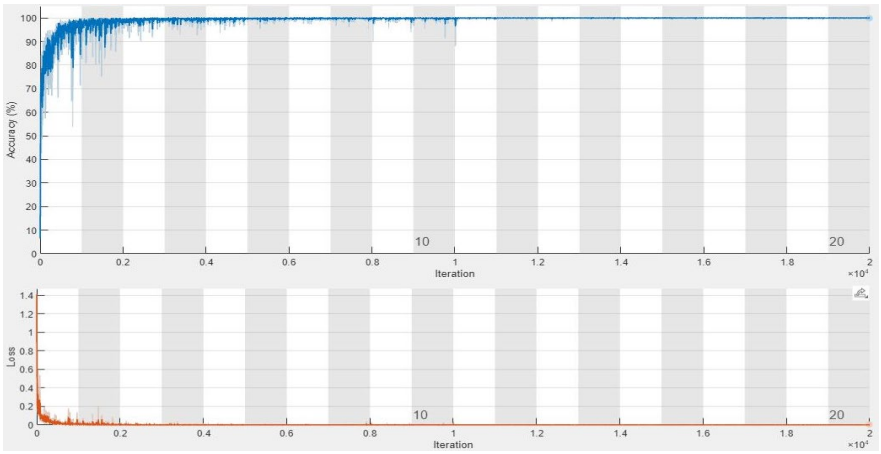


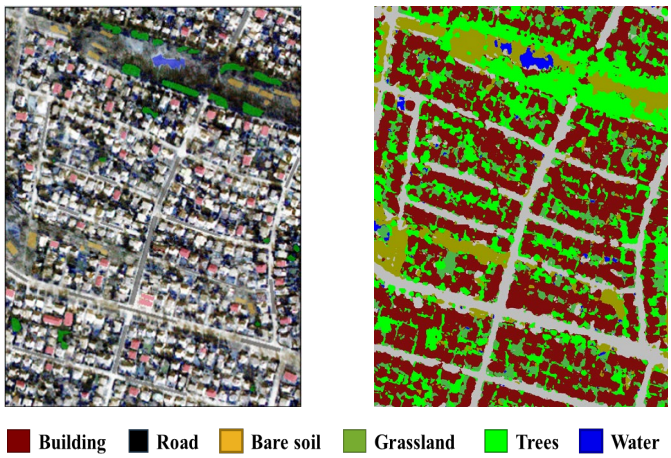
Figure 6
Loss versus accuracy evolution during the training progress

5 Results and Discussions

5.1. Classification Outcomes

The semantic segmentation using CNN requires representing training areas for effective image classification. The appropriate training samples are selected by visualizing the multi-spectral image in MATLAB Image Labeler where areas of interest are selected for each class linked to the real-world land cover (Figure 7a). The suggested CNN model applies the semantic segmentation for the multi-spectral image integrated with DSM into six main classes: water, grasslands, trees, bare land, roads, and buildings (Figure 7b).

Convolutional neural networks provide promising results for urban features such as roads and buildings [29]. The input image is classified into the main six urban classes; then, building and road features are detected (Figure 8a). The extracted features have been refined using the elevation values of the DSM to remove the confusion between urban and non-urban features using a certain threshold (Figure 8b). Reference buildings have been obtained in a vector format from Geofabric, free geodata created by projects like OpenStreetMap (OSM) [30]; meanwhile, road polygons were produced manually in QGIS using the OSM base map with minimal effort since there aren't many roads in the study area (Figure 8c).



(a) Training areas (b) Classified image

Figure 7
Training areas, classified image using MS image and DSM

(a) Classified features (b) Refined features (c) Reference features

Figure 8
Extracted, refined and reference urban features: buildings and roads

5.2. Building Extraction Evaluation

Building extraction outcomes are assessed using four accuracy indices: completeness, correctness, quality and overall accuracy (Eqs. 1-4). By comparing the extracted pixels with reference data, the True Positives (TP), False Positives (FP), and False Negatives (FN) are estimated for buildings while True negative (TN) is calculated numerically. Among all the extracted objects, TP represents the number of correctly classified pixels; FP represents the incorrectly classified pixels, while among the non-extracted or missed targeted objects, FN represents the non-extracted building pixels; TN represents the number of non-building pixels that are correctly classified [31].

$$\text{Completeness} = \text{TP}/(\text{TP}+\text{FN}) \quad (1)$$

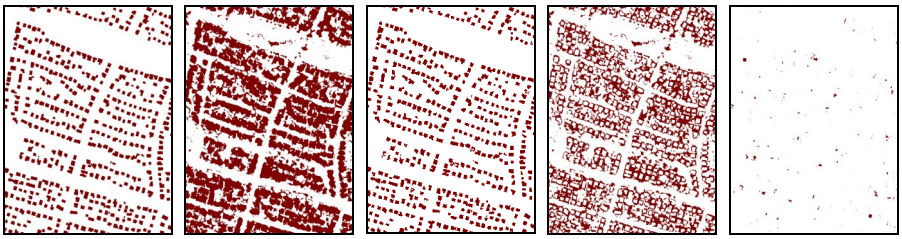
$$\text{Correctness} = \text{TP}/(\text{TP}+\text{FP}) \quad (2)$$

$$\text{Quality} = \text{TP}/(\text{TP}+\text{FP}+\text{FN}) \quad (3)$$

$$\text{Overall accuracy} = (\text{TP}+\text{TN})/(\text{TP}+\text{FP}+\text{FN}+\text{TN}) \quad (4)$$

5.2.1. Building Classification Results

Using the reference building pixel number (439670) and the classified building pixel number (1218700), the correctly classified building pixels are (TP 424867) where the classification matches the reference, the incorrectly classified building pixels are (FP 793831) where the classification does not meet the reference. The building pixels that are classified as non-building (FN 14803), and the non-building that are correctly classified are determined to be (TN 1387550) (Fig. 9).



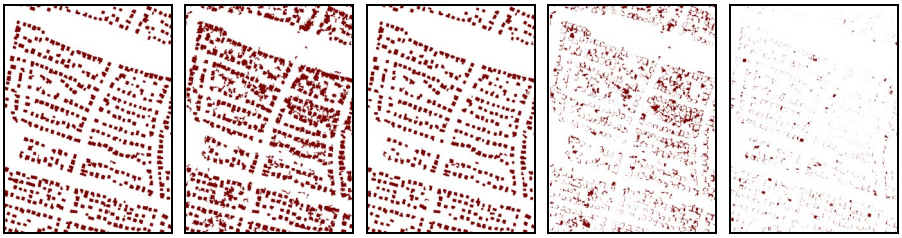
Reference pixels Classified pixels TP FP FN

Figure 9
Reference and classified pixels for buildings

5.2.2. Building Refinement Results

The accuracy indices are applied to the refined image, like the classified one, to show the effectiveness of using DSM in building extraction enhancement. The building reference pixels are the same (439670). The refined pixels are (618629), where (TP 394861) of them are correctly matching to the reference, and

(FP 223768) have no link with the real-world buildings. The building pixels that are misclassified as non-building are (FN 44809), and the non-building that are classified correctly are counted to be (TN 1987613) (Figure 10).



Reference pixels Refined pixels TP FP FN

Figure 10
Reference and refined pixels for buildings

The accuracy indices are calculated to quantitatively evaluate the classification and refinement processes. The classified and refined images have achieved a completeness of 96.6% and 89.8%, a correctness of 34.9% and 63.8%, a quality of 34.4% and 59.5%, and an overall accuracy of 69.1% and 89.9% respectively (Figure 11). Numerically, a considerable enhancement is notable for correctness, quality, and overall accuracy, while a light drop occurred in the completeness as the refinement using DSM has removed some buildin pixels from the scene.

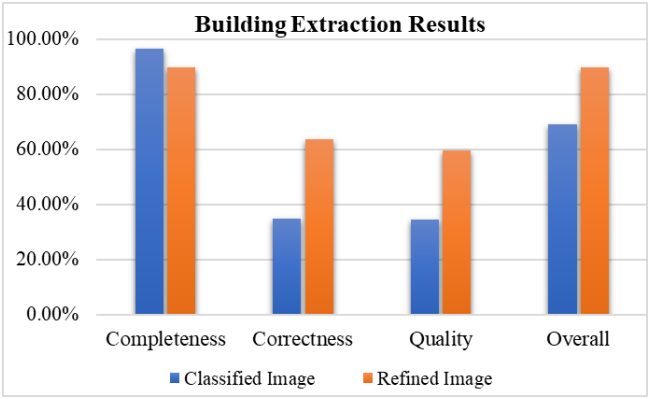


Figure 11
The accuracy indices for the classified and refined buildings

Qualitatively assessment is applied to the classified and refined outcomes and check the effect of the DSM refinement on the extracted buildings. The reference, classified, and refined images are vectorized to highlight the borders of each building (Figure 12). Intersection strategy is implemented to illustrate the matching and differences of the object borders. A promising correlation arises

between the refined and reference features which eliminate the misclassified regions of the classified image (Figure 13).

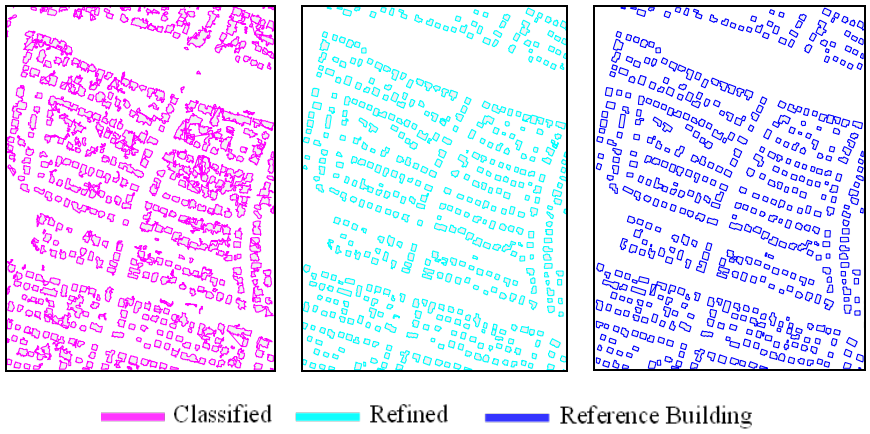


Figure 12
Extracted, refined and reference building borders

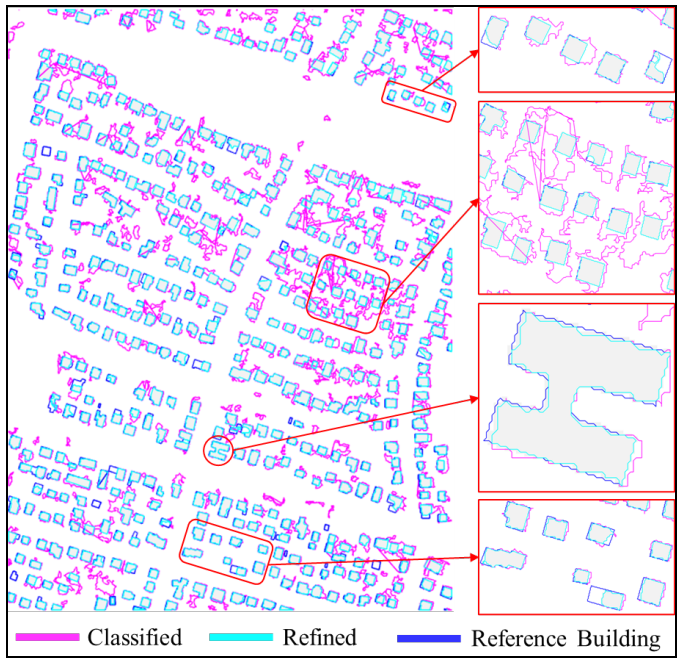


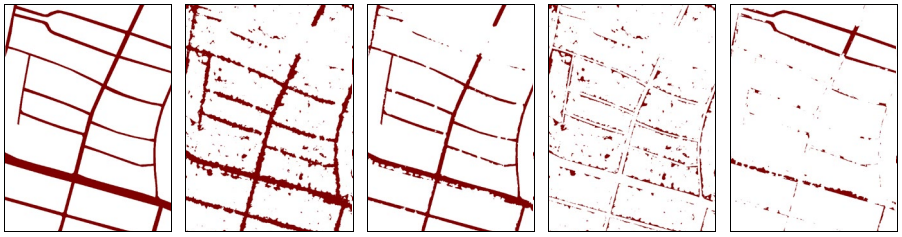
Figure 13
The matching between the extracted, refined and reference buildings

5.3. Road Extraction Evaluation

Road extraction assessment is applied focusing on the area of the extracted roads by calculating the number of road pixels for reference, extracted and refined objects. Evaluating road pixels is an appropriate method for calculating the accuracy indices. Calculating the extracted road areas is mainly affected by the variability of the road width. As a result, non-uniform width or non-accurate width roads, due to tree cover or similarity between road and neighbouring sand soil, show misclassified pixels near the edges of the road. The main objective of the road classification process is to identify the road class among the different land cover classes. The spectral similarity between roads and buildings or roads and parking lots is the main challenge.

5.3.1. Road Classification Results

The reference road pixel number (364539) and the classified road pixel number (400396) are calculated. The correctly classified road pixels are (TP 250228) where the classification matches the reference, and the incorrectly classified road pixels are (FP 150168) where the classification does not meet the reference. The road pixels that are classified as non-road, e.g. the road segments covered by trees and classified as vegetation, are (FN 114311), and the non-road that are correctly classified are determined to be (TN 2106344) (Figure 14).

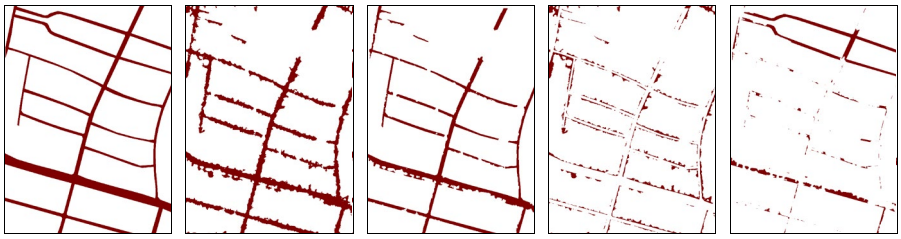


Reference pixels Classified pixels TP FP FN

Figure 14
Reference and classified pixels for roads

5.3.2. Road Refinement Results

Road refining is necessary to eliminate non-road regions, fill gaps, and trim road regions. The road reference pixels are the same (364539). The refined pixels are (372183), where (TP 246470) of them are correctly matching to the reference, and (FP 125713) have no link with the real-world roads. The road pixels that are misclassified as non-road are (FN 118069), and the non-road that are classified correctly are counted to be (TN 2130799) (Figure 15).



Reference pixels Refined pixels TP FP FN

Figure 15
Reference and refined pixels for roads

The classified and refined roads have achieved a completeness of 68.6% and 67.6%, a correctness of 62.5% and 66.2%, a quality of 48.6% and 50.3%, and an overall accuracy of 89.9% and 90.7% respectively (Figure 16). Numerically, using the DSM for road refinement adds a slight enhancement in the correctness, quality, and overall accuracy.

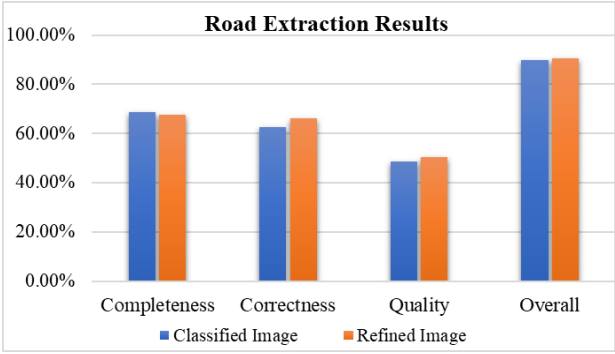


Figure 16
The accuracy indices for the classified and refined roads

Conclusions

Neural networks achieve significant progress in image classification and feature extraction tasks for sustainable urban applications using satellite images and DSMs. CNNs have the strength of high flexibility to different spatial and various spectral image characteristics. A U-NET model was proposed for image semantic segmentation using multi-spectral image, and extracted urban features have been refined using elevation values of DSM. Promising building and road features are experienced using U-net with several useful applications including monitoring urban expansion, real-time tracking, and disaster management. The obtained results evaluation, using accuracy metrics, demonstrates the effectiveness of the convolutional neural networks for image semantic segmentation of multimodal data and the efficacy of DSM to enhance the extratced features.

Acknowledgements

The research reported in this paper is part of project no. BME-NVA-02, implemented with the support provided by the Ministry of Innovation and Technology of Hungary from the National Research, Development, and Innovation Fund, financed under the TKP2021 funding scheme.

References

[1] T. H. Son, Z. Weedon, T. Yigitcanlar, T. Sanchez, J. M. Corchado, and R. Mehmood, "Algorithmic urban planning for smart and sustainable development: Systematic review of the literature," *Sustainable Cities and Society,* p. 104562, 2023

[2] F. Li, T. Yigitcanlar, M. Nepal, K. Nguyen, and F. Dur, "Machine learning and remote sensing integration for leveraging urban sustainability: A review and framework," *Sustainable Cities and Society,* p. 104653, 2023

[3] H. A. Al-Najjar *et al.*, "Land cover classification from fused DSM and UAV images using convolutional neural networks," *Remote Sensing,* Vol. 11, No. 12, p. 1461, 2019

[4] F. Jahan, J. Zhou, M. Awrangjeb, and Y. Gao, "Fusion of hyperspectral and LiDAR data using discriminant correlation analysis for land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* Vol. 11, No. 10, pp. 3905-3917, 2018

[5] T. Kwak and Y. Kim, "Semi-supervised land cover classification of remote sensing imagery using CycleGAN and EfficientNet," *KSCE Journal of Civil Engineering,* Vol. 27, No. 4, pp. 1760-1773, 2023

[6] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS journal of photogrammetry and remote sensing,* Vol. 152, pp. 166-177, 2019

[7] B. Neupane, T. Horanont, and J. Aryal, "Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis," *Remote Sensing,* Vol. 13, No. 4, p. 808, 2021

[8] F. Bore and A. Taraldsen, "Deep Convolutional Neural Networks for Semantic Segmentation of Multi-Band Satellite Images," Master's thesis Information- and communication technology - University of Agder, 2018 [Online] Available: https://uia.brage.unit.no/uia-xmlui/handle/11250/2563316

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 2015: Springer, pp. 234-241

[10] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems,* Vol. 25, 2012

[12] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700-4708

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778

[15] N. Mboga, C. Persello, J. R. Bergado, and A. Stein, "Detection of informal settlements from VHR images using convolutional neural networks," *Remote sensing,* Vol. 9, No. 11, p. 1106, 2017

[16] M. Fawzy, G. Szabó, and A. Barsi, "A Shallow Neural Network Model for Urban Land Cover Classification Using VHR Satellite Image Features," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* Vol. 10, pp. 57-64, 2023

[17] M. Fawzy, "Urban Feature Extraction From High Resolution Satellite Images.," Civil Engineering Dept., South Valley University, Faculty of Engineering, 2020

[18] B. Huang, B. Zhao, and Y. Song, "Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery," *Remote Sensing of Environment,* Vol. 214, pp. 73-86, 2018

[19] M. Fawzy, Y. G. Mostafa, and F. Khodary, "Automatic Indices Based Classification Method for Map Updating Using VHR Satellite Images," *JES. Journal of Engineering Sciences,* Vol. 48, No. 5, pp. 845-868, 2020

[20] E. S. A. ESA. "WorldView-2 European Cities." https://earth.esa.int/eogateway/catalog/worldview-2-european-cities (accessed 08 June, 2024)

[21] E. S. A. ESA. "WorldView-2 Instruments." https://earth.esa.int/eogateway/missions/worldview-2 (accessed 08 June, 2024)

[22] Q. Zhou, "Digital elevation model and digital surface model," *International Encyclopedia of Geography: People, the Earth, Environment and Technology,* pp. 1-17, 2017

[23] OpenDEM. "Free Digital Elevation Models." https://opendem.info/link_dem.html (accessed 08 June, 2024)

[24] S. Gehrke, K. Morin, M. Downey, N. Boehrer, and T. Fuchs, "Semi-global matching: An alternative to LIDAR for DSM generation," in *Proceedings of the 2010 Canadian Geomatics Conference and Symposium of Commission I*, 2010, Vol. 2, No. 6

[25] Geoportal Berlin. "airborne laser-scanned data - Berlin." https://fbinter.stadt-berlin.de/fb/ (accessed 08 June, 2024)

[26] ERDAS, *ERDAS Field Guide*. Erdas, 2021

[27] C. Pohl and J. Van Genderen, *Remote sensing image fusion: A practical guide*. Crc Press, 2016

[28] M. Fawzy and A. Barsi, "A U-Net Model for Urban Land Cover Classification Using VHR Satellite Images," *Periodica Polytechnica Civil Engineering,* 2024, doi: https://doi.org/10.3311/PPci.37599

[29] C. Ayala, R. Sesma, C. Aranda, and M. Galar, "A deep learning approach to an enhanced building footprint and road detection in high-resolution satellite imagery," *Remote Sensing,* Vol. 13, No. 16, p. 3135, 2021

[30] Geofabrik. "OpenStreetMap Data Extracts." https://download.geofabrik.de/ (accessed 20 May, 2025)

[31] A. Shukla and K. Jain, "Automatic extraction of urban land information from unmanned aerial vehicle (UAV) data," *Earth Science Informatics,* Vol. 13, No. 4, pp. 1225-1236, 2020