Received: 10 March 2026 | Revised: 9 June 2026 | Accepted: 3 September 2026 | Published online: 20 October 2026

RESEARCH ARTICLE

Journal of Computational and Cognitive Engineering
2026, Vol. 00(00) 1–16
DOI: 10.47852/bonviewJCCE62029582

BON VIEW PUBLISHING

Beyond Complexity: Evaluating Simple vs. Complex Deep Learning Architectures for Urban Density Mapping

Willy Boen[1] **and Edy Irwansyah**[1,*]

[1] *School of Computer Science, Bina Nusantara University, Indonesia*

Abstract: The spatial distribution of population density and land use is a fundamental metric for targeting spatial planning policies under the Sustainable Development Goals. This study presents a comparative evaluation of deep learning architectures for urban density estimation, with an emphasis on spatial generalization. The study was conducted in Jakarta and its surrounding cities (Jabodetabek), a highly heterogeneous region that provides a rigorous setting for evaluating model generalization. Sentinel-2 satellite imagery and European Space Agency WorldCover 2020 data were acquired through Google Earth Engine, together with the Normalized Difference Vegetation Index and the Normalized Difference Built-up Index to explicitly represent vegetation and built-up characteristics. Quantitative evaluation showed that all deep learning models performed well, with R^2 exceeding 0.94 in all experiments. The U-Net architecture achieved the highest predictive accuracy, with the lowest mean squared error (MSE = 0.0006954). Notably, the lightweight convolutional neural network (CNN) proved to be a highly competitive alternative (MSE = 0.0006998), requiring approximately four times fewer parameters and delivering roughly twice the inference speed of the deeper U-Net and ResNet models. In contrast, complex models such as EfficientNet and ResNet showed diminishing returns in accuracy relative to their design complexity, while classical machine learning baselines (Random Forest and XGBoost) failed to capture spatial configurations robustly. These findings demonstrate that increasing network depth does not necessarily improve performance in spatial fractional regression. Therefore, model selection should balance predictive accuracy against computational and environmental costs.

Keywords: deep learning, urban density, Jakarta city, Sentinel-2, mapping

1. Introduction

Spatial information on population density and land use is a crucial instrument for supporting targeted spatial planning policies under the framework of Sustainable Development Goals (SDGs). This is because accurate density data enable policymakers to predict future land consumption to ensure the sustainability of urban land resources. Recently, density estimation has shifted from conventional statistical methods, such as linear regression, to the use of machine learning (ML) models, such as Random Forest (RF), which have proven more effective at capturing connectivity and density variations in new development areas [1].

In heterogeneous metropolitan regions, population distribution is found to be far more influenced by socioeconomic features, such as point-of-interest (POI) density and building volume, than by natural environmental factors. The utilization of ensemble learning techniques through stacking models has been proven to be capable of producing high-precision population spatialization at the pixel scale in dense metropolitan areas [2]. Accordingly, deep learning (DL) has become the primary trend in the semantic segmentation of urban features, where encoder–decoder architectures, such as U-Net and Attention mechanisms, have been identified as the most effective methods for extracting building objects and infrastructure [3].

The implementation of the U-Net model has been validated in various contexts, including the mapping of Urban Green Spaces (UGS) at the metropolitan level, with an overall accuracy (OA) of 97% using very-high-resolution satellite imagery [4]. The superiority of the U-Net architecture is also evident in its ability to segment cityscape elements with an average accuracy of 96.4%, which is critical for the development of robotic navigation systems and visual urban environment understanding [5]. However, a primary challenge in detailed urban-object analysis is the limited resolution of freely available satellite imagery [6].

In particular, the optimal use of DL-based super-resolution frameworks can improve the definition of buildings in medium-resolution imagery [6]. As a result, optical and synthetic aperture radar data fusion strategies have also demonstrated higher precision in mapping building footprints as well as road networks than using a single type of sensor [7]. In addition, OpenStreetMap (OSM)-based automated data such as w/o trained DL on a global scale can be used to generate labels directly for training [8], which improves mapping efficiency of impervious surfaces in the infrastructure.

***Corresponding author:** Edy Irwansyah, School of Computer Science, Bina Nusantara University, Indonesia. Email: edirwan@binus.ac.id

© The Author(s) 2026. Published by BON VIEW PUBLISHING PTE. LTD. This is an open access article under the CC BY License (https://creativecommons.org/licenses/by/4.0/).

In rapidly urbanizing regions, the use of aerial photography from drones allows for the automated estimation of the building coverage ratio (BCR) with an accuracy exceeding 94% [9]. Additionally, unsupervised DL methods have been developed to generalize urban forms without dependence on manual labeling and are capable of consistently describing the physical distribution of buildings [10]. Dynamic monitoring of infrastructure growth can also be performed using transfer learning (TL) techniques to detect urban changes based on a comparison of segmentation masks over time [11].

Architectural innovation continues to evolve with the emergence of transformer-based models, such as Swin-U-Net, which demonstrate superiority in capturing complex spatial details compared with conventional convolutional neural network (CNN) models [12]. Although labeled data are often limited, fine-tuning strategies on pretrained models have proven effective in increasing land-use classification accuracy [13]. Furthermore, density modeling has been integrated with city vitality predictions to understand the continuous interaction between physical morphology and human movement flows [14]. Theoretically, these density estimations are supported by flexible generative frameworks for modeling various types of complex data [15].

Most prior research trained on one task type (for instance, image classification or object detection) as a single architecture and did not compare modern models like EfficientNet, ResNet, or U-Net concurrently across broader density estimation tasks. In addition, the problem of spatial generalization from one geographic region to another is still one of the largest barriers to adopting DL models in remote sensing practice. Thus, the objective of this research is to understand the generalizability in urban density estimations of various model architectures that could provide a more accurate and robust prediction throughout heterogeneous regions such as the Jabodetabek area in Indonesia.

2. Literature Review

The development of a roundtrip framework for density estimation based on deep generative neural networks has successfully combined the generative power of models, such as generative adversarial networks (GANs), with the ability to provide accurate density value estimations. This model is capable of general-purpose density estimation without requiring strict architectural constraints, making it more flexible for various types of complex data than previous neural density estimators [15].

A systematic review and meta-analysis of various studies on the semantic segmentation of urban features indicated that encoder–decoder architectures, such as U-Net and Attention mechanisms, represent the most effective primary trends. The greatest challenges identified in this analysis are the heterogeneity of satellite data and the scarcity of well-annotated training data for detecting elements such as buildings, roads, vegetation, and water [3].

Mapping UGS at the metropolitan level using very-high-resolution satellite imagery and DL techniques for semantic segmentation has achieved an OA of 97%. This study demonstrated that the U-Net architecture with various encoders is capable of mapping and categorizing UGS into five specific types: road green belts, residential parks, roundabouts, squares, and city parks [4].

The detection of building footprints and roads in high-resolution satellite imagery can be enhanced by modifying the U-Net architecture using a ResNet-34 backbone. The integration of optical data from Sentinel-2 and radar data from Sentinel-1 has proven to be superior in producing maps with a 2.5 m (sub-pixel) resolution, which is four times more detailed than the native resolution of the input imagery [7].

Automated detection of impervious surfaces using DL can be performed efficiently by leveraging OSM data as automated labels to generate large-scale training datasets. The U-Net model, combined with satellite band features and statistical indices, yielded the best results, with an accuracy of 92.46% in predicting impervious surface fractions across various terrains [8].

The BCR in rapidly urbanizing areas can be estimated using the ResNet101-based Mask R-CNN model. BCR estimates from drone-acquired aerial photos can classify structures over extensive regions in a relatively short time frame, with median classification accuracy exceeding 94% at the block scale and 99% at the parcel scale [9].

Although DL is a possible contender, the general breakthroughs in the overall ML field suggest that there are many nonlinear families that can achieve high-accuracy modeling very well. Relatively simple architectures of recurrent neural networks can be used with great accuracy to predict high-frequency price indices [16]. Moreover, research on agricultural commodities indicates that nonlinear autoregressive neural networks have the capability to learn complex dynamics while providing stable low-level prediction error rates [17].

Apart from neural networks, there are other advanced statistical learning approaches that compete well regarding the capacity to model complex patterns. Bayesian optimization has shown promise and effectiveness in the daily prediction of regional commodity price index using Gaussian process regression models [18]. The Vector Error-Correction Modeling combined with directed acyclic graphs, which are suited for analyzing spatial interrelationships, combines rigorous criteria for determining contemporaneous causal direction in regional markets [19]. Further, composite forecasting and ensemble models based on a number of trimming strategies have been pushed as a way to diversify against model misspecifications and outperform across horizons [20, 21].

A new unsupervised DL method, the Multiscale Residual Convolutional Autoencoder, was developed to generalize urban forms without depending on manual labels. This study successfully identified urban form clusters capable of explaining much of the height and area distribution of buildings with very low image reconstruction error rates [10].

High-precision population spatialization in metropolitan regions can be achieved using the GXLS-Stacking model, which integrates the gradient boosting decision tree (GBDT), XGBoost, LightGBM, and support vector regression (SVR) algorithms. Analytical results show that socioeconomic features, such as POIs and building volume, exert a far more dominant influence (99.54%) on population distribution than natural environmental features [2].

Urban change detection using aerial imagery can be optimized using the U-Net architecture with a ResNet-101 backbone and TL techniques. This methodology is effective for identifying both new and demolished buildings by automatically comparing segmentation masks from different time periods [11].

Earlier investigations also confirmed that RF-based models are better at estimating differences in urban population density than using traditional linear regression. Of the many factors studied in population density estimation, new development area density variations and regional connectivity level are the most significant ones [1].

The comparison of different super-resolution frameworks leveraging DL approaches with respect to Sentinel-2 imagery

showed that GAN-based methods like the enhanced super-resolution generative adversarial network (ESRGAN) yield the sharpest visual textures. Nevertheless, CNN-based architectures such as the residual channel attention network (RCAN) and the enhanced deep super-resolution network (EDSR) give more stable quantitative accuracy for better resolution improvement of building and infrastructure shapes in dense urban areas [6].

Research has shown the suitability and effective learning of transformer-based architectures, for example, the Swin-U-Net to capture complicated spatial details in high-resolution imagery for land-use classification. With regard to dense pixel-wise classification, the performance of Swin-U-Net significantly outperforms conventional CNN-based models, that is, the fully convolutional network (FCN) or SegNet; their respective overall accuracies achieved by using basic features from one source change by approximately 50.0% on the Australian benchmark across different urban-object categories in Cityscapes [12].

The U-Net architecture was further confirmed by performing semantic segmentation of urban elements in the Cityscapes dataset, capturing complex aspects like as roadways, walkways, and vegetation. The average accuracy of this model was 96.4%; therefore, it is proper to use it as a reliable solution for applications of robotic navigation and urban environment scene understanding from street-level vision [5].

The fine-tuning of Inception-v3 using TL is particularly powerful for land-use classification, even with few labeled samples. This allows models to use learnt features from ImageNet and other large datasets and obtain state-of-the-art outcomes in summarization for urban land scene categories [13].

The Urban Vitality Prediction Network model, which innovatively includes squeeze-and-excitation (SE) blocks and residual connection with coordinate attention (RCA) bottlenecks, is capable of predicting urban vitality at the regional scale. This approach significantly decreases the prediction errors for the static population density and dynamic pedestrian flows, assuming road networks are the main factors that determine population density [14].

In summary, although numerous research have shown the great potential of DL for urban feature mapping from high-resolution satellite images, there are some fundamental limitations on which this study is built. First, conventional ML and classical statistical approaches are not well suited to model the nonlinear spatial interactions and extremely varied population distribution patterns that take place inside metropolitan regions.

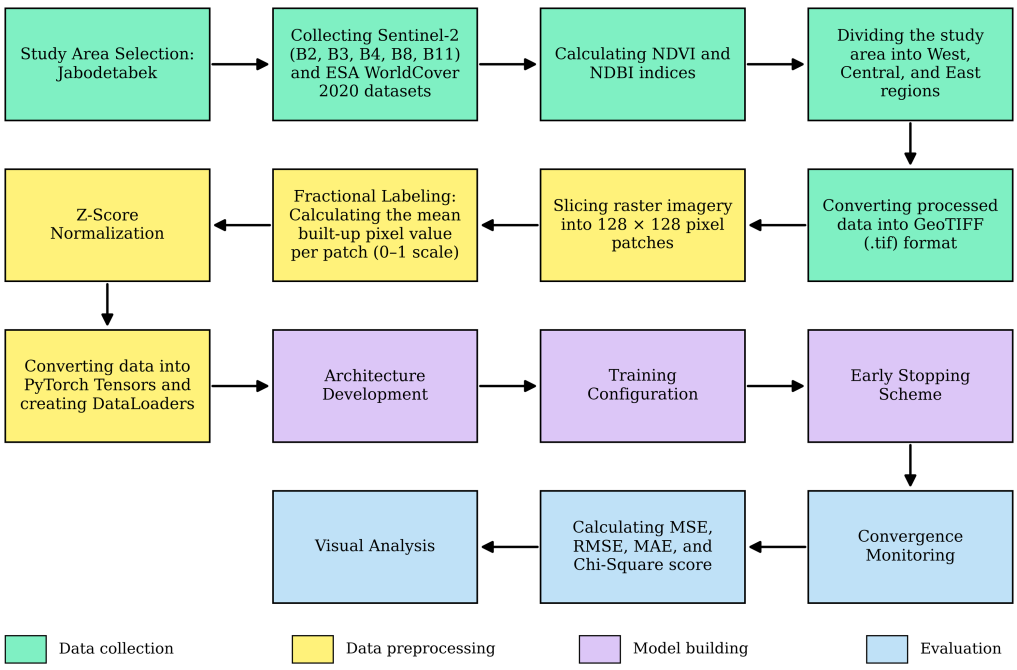
Second, previous studies have mainly used a single model architecture for specialized classification or segmentation tasks (*e.g.*, green spaces vs. building footprints) and have therefore been comparatively limited by not trying to produce code by evaluating similar contemporary models like EfficientNet, ResNet, and U-Net as well across broader density-estimating goals.

This study thus rationalizes the choice of DL, not just as a simple default selection but as an informed decision over other strong (nonlinear) modeling techniques such as Gaussian process regression or ensemble methods. In this comparison, the unique ability of CNNs to perform feature extraction is specifically weighed against the efficiencies achieved by conventional non-DL ML models that can also handle spatial data.

3. Research Methodology

The workflow of our study is shown in Figure 1, and each color corresponds to a step in our research procedure. The steps involved were data collection, dataset preprocessing, model building, and evaluation process. First, we collected and preprocessed the relevant data to ensure a suitably high-quality dataset. Next, modeling was done on four models with various architectures: CNN, EfficientNet, ResNet, and U-Net. The mean squared error (MSE) training and validation loss curves were used to evaluate these models. In addition, the models were evaluated on the test dataset and then analyzed using regression plots between predictions and ground truth. Lastly, residuals and absolute residuals vs. actual values were examined.

Figure 1
Research workflow



3.1. Data collection

The area of interest for the defined study is Jakarta and its respective follow-agency cities (Bogor, Depok, Tangerang and Bekasi – Jabodetabek) as presented in Figure 2, where the red, green, and blue blocks denote the western, central, and eastern partitions used for training, validation, and testing, respectively. This region was chosen for its high heterogeneity, which we expect will enable the models to generalize well. Data from the Sentinel-2 satellite and European Space Agency (ESA) WorldCover were retrieved through the Google Earth Engine (GEE).

Sentinel-2 is an extensive multispectral instrument with sensors primarily used in Earth surface observation, maintaining the capacity for high resolution, wide coverage, and high geometric performance [22]. ESA WorldCover is a global land cover map from 2020 at 10 m resolution, based on Sentinel-1 and Sentinel-2 data [23]. Both of the datasets were collected in 2020. A close-up view of both satellite datasets is shown in Figure 3. For ESA WorldCover, the redder areas represent population/built-up density.

The present study considers fractional regression—a statistical technique that is specifically designed to analyze dependent variables in the form of fraction/proportion/percentage and lies within the limits of [0,1]. You are trained on 0 output: Your input was primary Sentinel-2 data, and the reference (ground truth) was ESA WorldCover. We used five Sentinel-2 bands, which were B2 (blue), B3 (green), B4 (red), B8 (near infrared), and B11 (short-wave infrared 1).

Nevertheless, these raw bands are by themselves not entirely sufficient to capture the land cover characteristics optimally as models typically benefit from having derivative features that represent certain physical phenomena explicitly. Thus, spectral indices—Normalized Difference Vegetation Index (NDVI) and Normalized Difference Built-up Index (NDBI)—were added as auxiliary features in this study (see Figure 4).

Reliable NDVI time series data are key in supporting vegetation condition and providing insights into the relationship between environmental forcing and the status of vegetation growth, productivity, biodiversity, and plant health [23, 24]. The NDBI is analogous to the NDVI and was developed to highlight built-up areas, exploiting the higher reflectance of built-up surfaces in the short-wave infrared (SWIR) band relative to the near-infrared (NIR) band [25]. Based on NDVI and NDBI density results, a significant loss of vegetation land and a substantial gain in urbanization (and built-up) land use were reported [26]. Mathematically, they are formulated as:

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{1}$$

$$NDBI = \frac{SWIR - NIR}{SWIR + NIR} \tag{2}$$

Since deriving the data from GEE must not interfere with the experimental design, some preprocessing steps were performed in this instance. In order to reduce the presence of atmospheric artifacts, a cloud cover threshold lower than 10% was applied to the Sentinel-2 collection, and invalid pixels were corrected based on a median composite method over time to obtain a clean spectral signature throughout the whole study area. The data were then split into three geographical regions: west, central, and east (Figure 2).

Figure 2
Jakarta and its surrounding cities (Bogor, Depok, Tangerang and Bekasi)—Jabodetabek, Indonesia

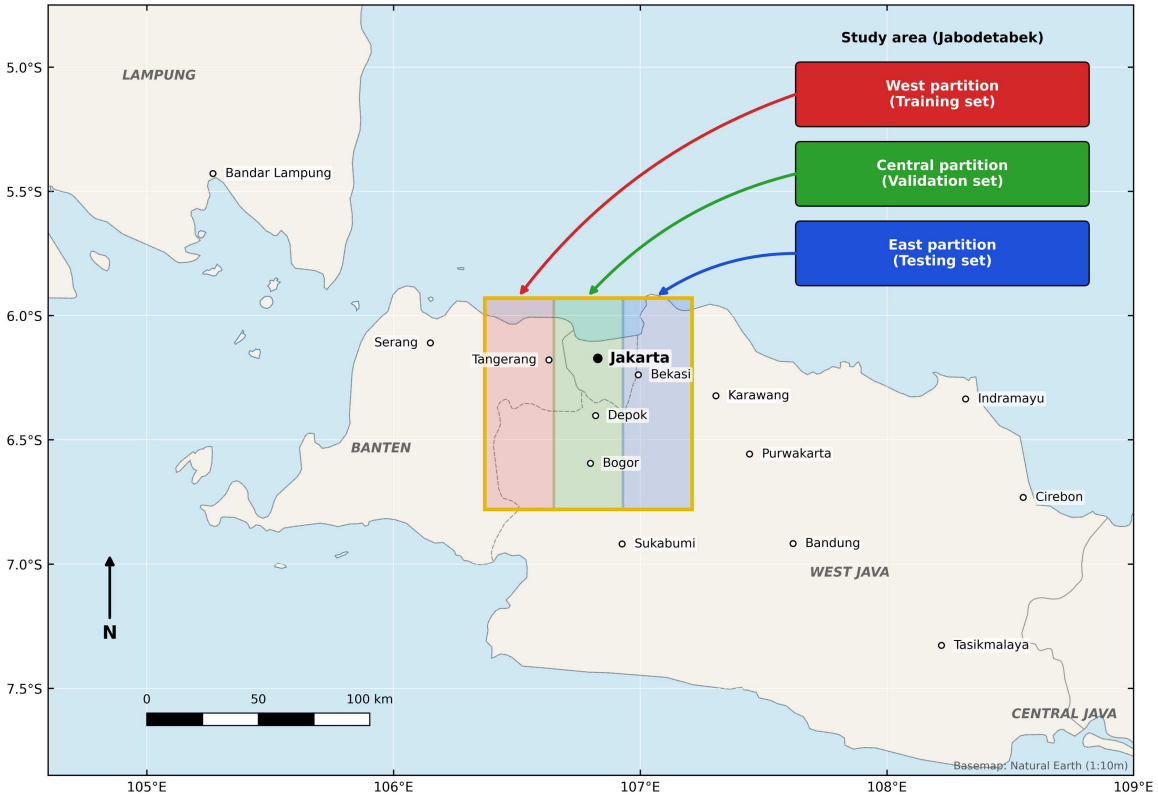


Figure 3
(a) Sentinel-2 satellite imagery and (b) built-up area of Jakarta city

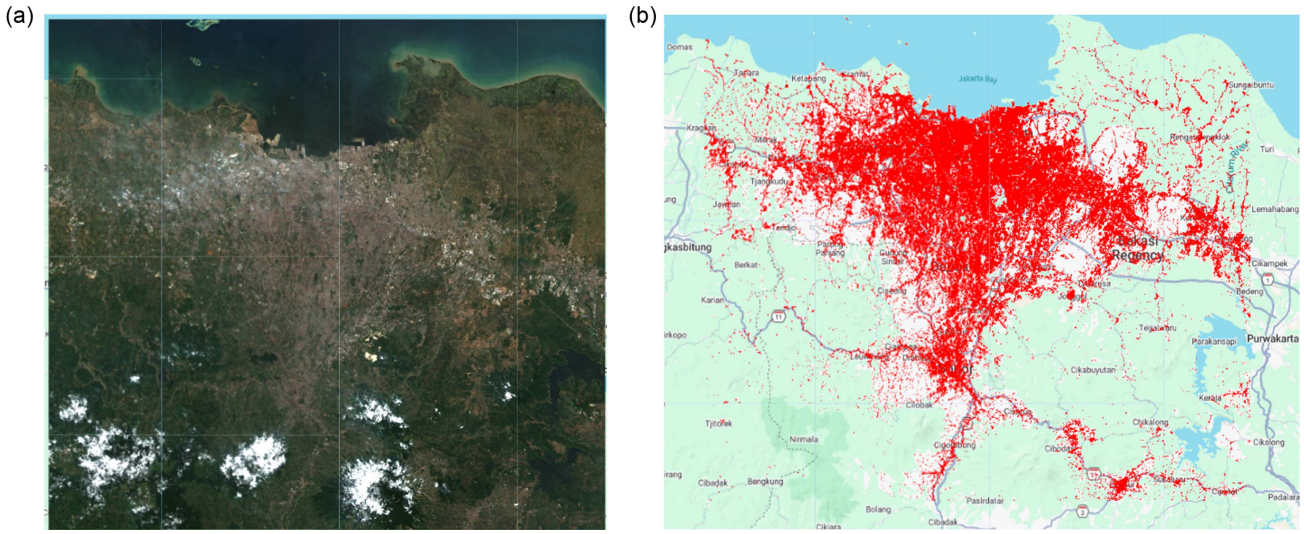
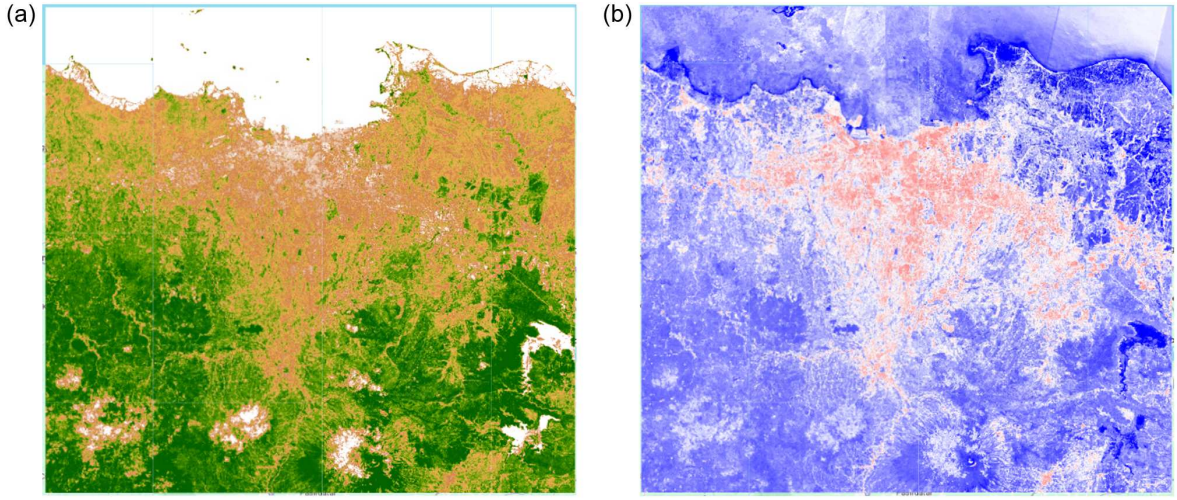


Figure 4
Normalized difference indices: (a) NDVI and (b) NDBI



As a matter of fact, spatial autocorrelation represents a serious problem that we solved by keeping a certain buffer zone between the western (training), central (validation), and eastern (testing) partitions. This was done to prevent adjacent patches from having the same spectral information due to spatial continuity by masking out pixels located within a predetermined distance from the borders. Finally, the resulting eight-band image (B2, B3, B4, B8, B11, NDVI, NDBI, and the built-up mask) was clipped to the three partitions and exported from GEE to Google Drive at 10 m resolution, from which three GeoTIFF (.tif) files (west, central, and east) were downloaded.

3.2. Data preprocessing

After data gathering, datasets were preprocessed. Three maps were categorized as training (west), validation (central), and testing (east) maps. For the patch extraction process, the raster images were extracted as patch-based samples of size 128 × 128 pixels. This patch size was chosen since it is large enough to maintain the structural texture and spatial context of Jabodetabek Urban Blocks, while preventing overhead in terms of DL architecture receptive fields. The extraction process was randomized to generate a specific number of discrete samples for each geographical region, yielding a total of 4,600 patches. This dataset size was precisely partitioned into 3,000 samples for the training set (west partition), 800 samples for the validation set (central partition), and 800 samples for the testing set (east partition).

The label for each extracted patch was computed as the average of all built-up pixels within that specific patch, which successfully yielded a target variable bounded between 0 and 1, mapping to the fraction of the built-up area. Furthermore, Z-score normalization of features was performed based on the mean and standard deviation calculated from the training set only to completely avoid data leakage. This statistical step involved clipping extreme outliers in the spectral data to significantly improve numerical stability and to avoid skewed normalization parameters. Finally, to utilize this dataset with the DL model, the arrays were converted into PyTorch tensors and stored securely within a DataLoader structure.

3.3. Modeling

At training time, the training data were shuffled to add some variation during learning from epoch to epoch while keeping valid and test data unshuffled so that evaluation was stable. Batching was done with a batch size of 32 samples per iteration.

3.3.1. CNN

Figure 5 shows that the model uses a CNN architecture for regression with a 7-channel multispectral input. It consists of four ConvBlocks, each containing two Conv2D layers (3 × 3 kernel, padding = 1), batch normalization, and rectified linear unit (ReLU). The first three blocks used 2 × 2 Max Pooling, whereas the fourth block omitted pooling to preserve the feature information. The number of feature channels increased from 32 to 64, 128, and 256 channels. Adaptive Average Pooling is then used to reduce the feature map to 1 × 1 per channel before flattening it into a feature vector. The final part was a fully connected layer (256–128–1) with ReLU and dropout (0.3).

3.3.2. EfficientNet

Figure 6 shows the model based on EfficientNet modified for regression. The first convolutional layer was adjusted for a 7-channel input. The backbone consists of MBConv blocks with ImageNet-pretrained weights. The default classification layer was replaced with an identity layer, and a regressor head consisting of dropout (0.3) and a fully connected layer (1280–1) was added to it.

3.3.3. ResNet

Based on ResNet in Figure 7, this architecture uses residual blocks with shortcut connections to maintain gradient stability. Each block consists of two 3 × 3 convolutions, followed by batch normalization and ReLU. The number of feature channels was increased from 64 to 512. After feature extraction, global average pooling (GAP) was applied, followed by a regressor head with two fully connected layers (512–128–1) and dropout (0.3).

3.3.4. U-Net

Specifically, the U-Net architecture has been modified from semantic segmentation to the regression-based urban density estimation task in this study. The structural modality of the modified architecture shown in Figure 8 inherits both the hierarchical feature extraction (i.*e*., encoder–decoder structure) and skip connections, thus denoted as UNetRegressor.

In this study, however, the standard classification head was swapped for a regression-based one. More specifically, the final output from the decoder is passed through a GAP layer to convert these multi-scale spatial features into a high-level feature representation and is lastly passed through some dense layers with linear activation. We used an adaptation that makes the model predict a single value (between 0 and 1), which corresponds to the estimated density, rather than a pixel-wise class map, which ensures fair comparison methodologically as well with other architectures based on regression since it can then be evaluated like a simple regression task for continuous distributions.

The training of all models was done using the MSE loss function and Adam optimizer. A ReduceLROnPlateau scheduler was employed to multiply the learning rate by 0.5 when the validation loss did not improve for five epochs in a row. A maximum of 200 epochs and early stopping if the validation loss did not improve within the last 20 cycles. For models with outputs that were spatially map-like (B 1 H W), a global mean was applied to produce a single prediction value per sample before MSE calculation.

In addition to the DL architectures, two nonlinear classical ML algorithms, like Random Forest Regressor and XGBoost Regressor, were integrated as baseline benchmarks to validate the necessity of deep representation learning. Unlike the convolutional networks that process 4D spatial tensors, these classical models ingest tabular features derived by computing the mean spectral values across all seven input channels for each individual patch. This structural

Figure 5
ConvBlock and CNN regression architecture

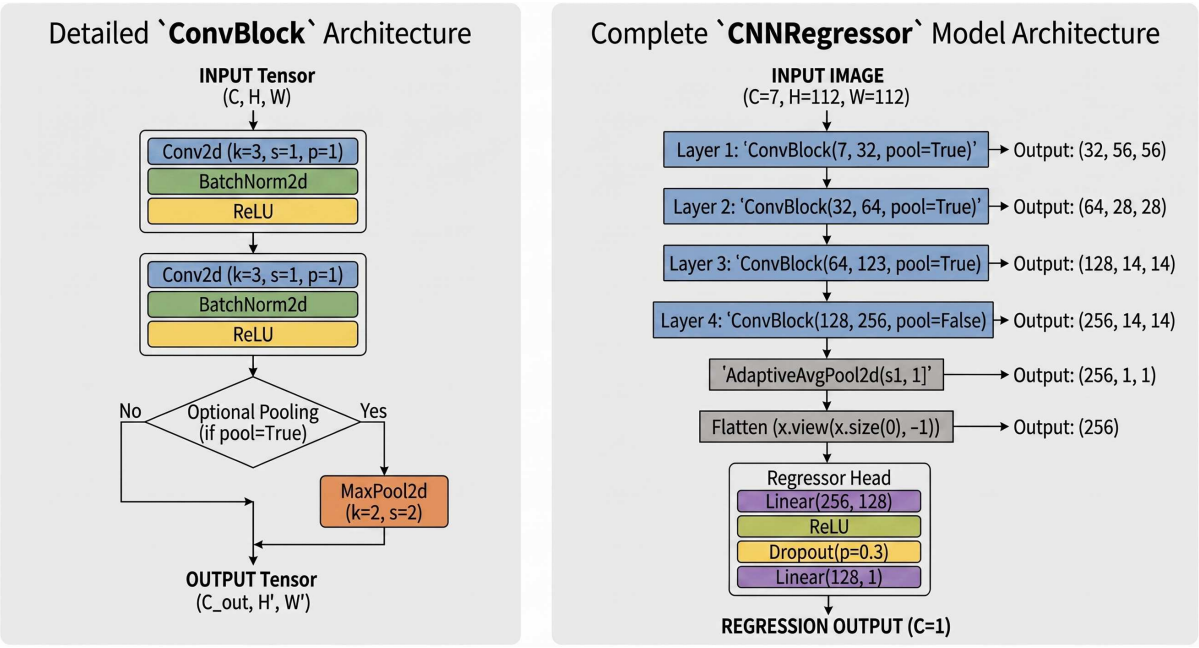


Figure 6
EfficientNet architecture

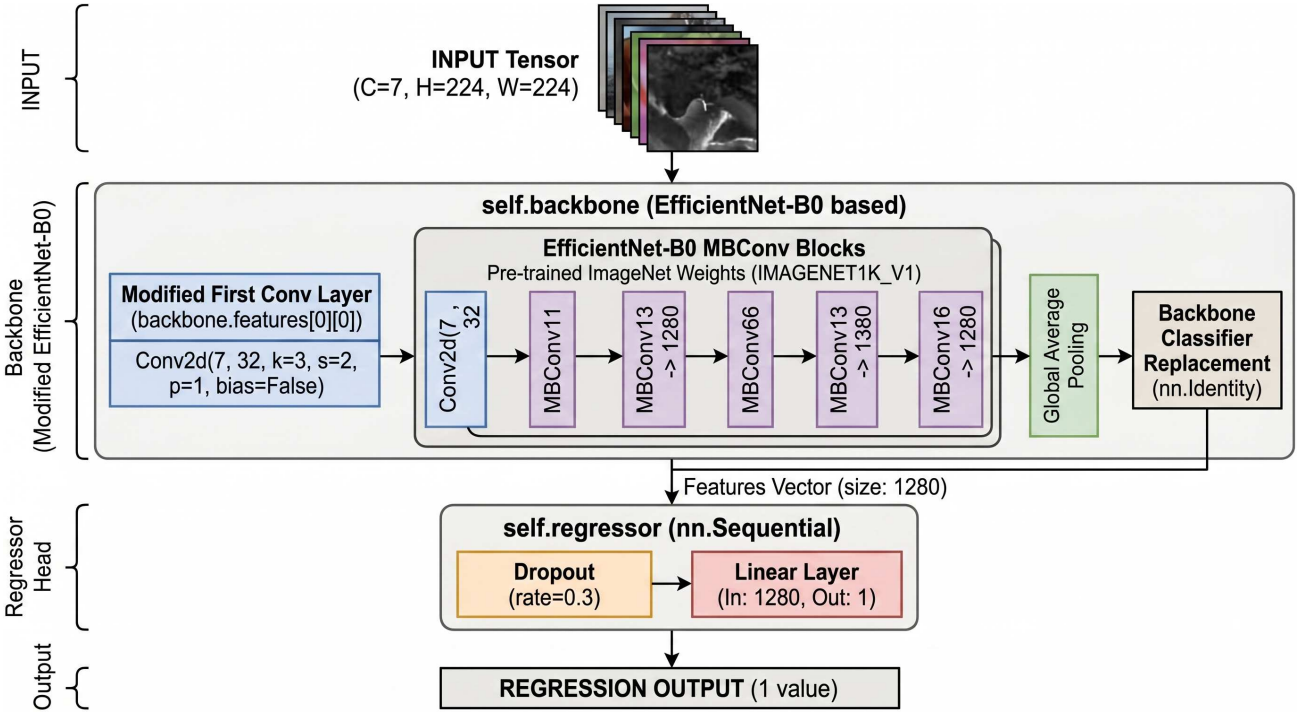
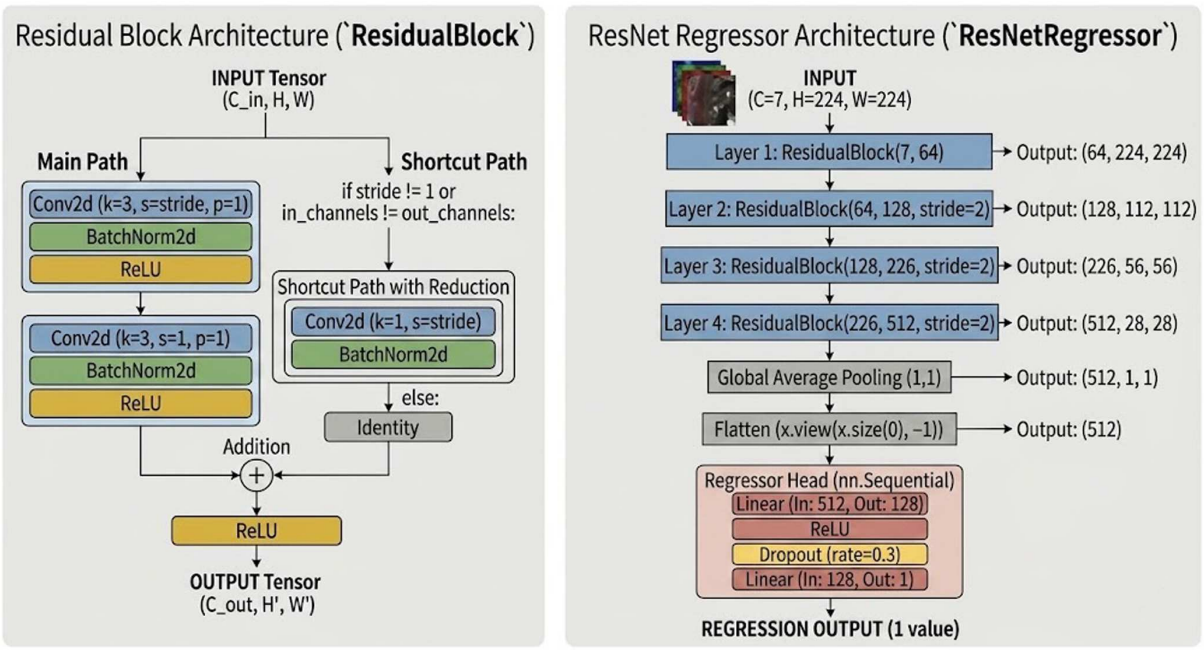


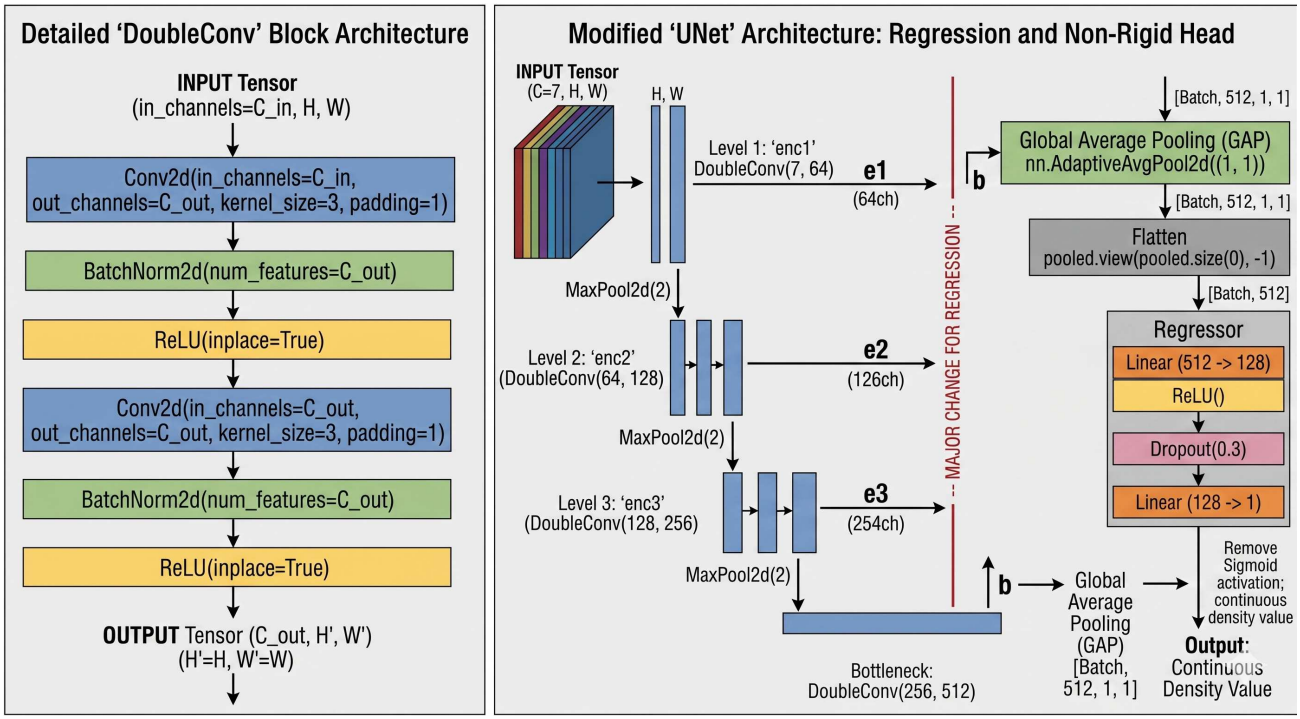
Figure 7
ResNet block and ResNet regression architecture



transformation reshapes the spatial dimensions into a distinct 2D matrix of 3,800 × 7 for the combined training pool and 800 × 7 for the testing set, ensuring an equitable evaluation baseline utilizing the exact same spatial sample partitions.

To guarantee full experimental reproducibility, all computational modeling was executed using Python (v3.11.14) and PyTorch (v2.11.0) frameworks. The hardware environment was accelerated by an NVIDIA GeForce GPU backed by a CUDA runtime library. During the optimization process, the initial learning rate for the Adam optimizer was strictly set to 10^{-4} with a weight decay coefficient of 10^{-4} to prevent overfitting. All architectural structures were initialized with a fixed random seed of 42 to ensure completely deterministic weight updates across all comparative runs.

Figure 8
U-Net architecture



3.4. Evaluation

The training and validation loss curves (MSE) were used to evaluate model performance, monitor convergence, and determine overfitting [27]. After that, we can evaluate models on the test set and print the following metrics: MSE, root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). Our exploratory data analysis was visual, using plots of prediction vs. real, residuals distribution, and absolute error against actual value. In addition, the relative root mean squared error (RRMSE) was computed to offer a normalized measure of error magnitude relative to the observed average density, enabling enhanced comparability across architectures and scales.

In the scope of evaluation, a computational cost analysis has been added to provide a fair assessment of the accuracy-complexity trade-off. It involved quantifying model complexity in terms of trainable parameters and Floating Point Operations (FLOPs), along with operational efficiency metrics like Total Training Time and Inference Time (ms/sample) evaluated on the same hardware configurations.

Finally, a paired t-test was performed to see if the changes in performance were statistically significant with respect to the absolute inaccuracy of the model predictions. Thus, this statistical rigor provides a clear answer to whether the efficiency gains of simpler architectures (*e.g.*, the CNN) involve a statistically significant sacrifice in terms of predictive accuracy or indeed provide a pragmatically appropriate solution for large-scale urban density mapping.

4. Result and Discussion

The experimental results and performance analysis of the proposed models in this study are presented in this section. We evaluated all architectures listed in Table 1 with the training, validation, and testing data. The performance of these models was quantitatively analyzed through MSE, RMSE, MAE, R^2, and the correlation coefficient (r). In addition, graphical analysis was carried out for comparing predictions with the actuals, distribution of residuals, and pattern of absolute errors, allowing a more thorough interpretation of the capacity of the model to predict target values.

Table 1
Model evaluation metrics against test data

Model	MSE	RMSE	MAE	R^2	Target Range	Baseline MSE
Random Forest	0.0100904	0.1004509	0.0592754	0.6481569	0 to 0.8621	0.0297924
XGBoost	0.0107155	0.1035156	0.0619109	0.6263605	0 to 0.8447	0.0297924
CNN	0.0006998	0.0264533	0.0150991	0.9748958	0 to 0.84936523	0.0278749
EfficientNet	0.0017097	0.0413490	0.0263083	0.9432946	0 to 0.8985596	0.0301513
ResNet	0.0010596	0.0325518	0.0189541	0.9683899	0 to 0.86639404	0.0335215
U-Net	0.0006954	0.0263702	0.0156074	0.9782967	0 to 0.857666	0.0320407

Figure 9 shows the training process and convergence tendency of four evaluated models (MSE curves). The first training epoch across all architectures had a high loss that dropped quickly, indicating its learned the basic structure of urban density from multispectral data quickly. In fact, the differences between training and validation curves were small for all models despite partitioning the dataset geographically into western, central, and eastern sections, which confirmed robust spatial generalization without major overfitting.

When we performed a comparative analysis of the convergence rates, we noticed different behaviors according to the architectural complexity. The CNN demonstrated a fairly consistent learning behavior, and MSE values converged and stabilized as early as around the 15th–20th epoch. Complex models like EfficientNet and ResNet showed huge shifts and very big spikes in their validation curve in the early epochs. Such discrepancies might be attributed to the weight normalization necessary for deeper networks; that is, they are more sensitive to the distribution of the validation data in early epochs. Both models suffer from pole instabilities at the beginning of training but reach zero loss after 10 epochs in the stable state.

The U-Net design has a fast learning rate; the loss for both training and validation drops rapidly, nearly immediately. The model converged around epoch 15, with the entire loss hitting convergence. It observed slight fluctuations in the mid-epochs, probably owing to variances in sample distribution.

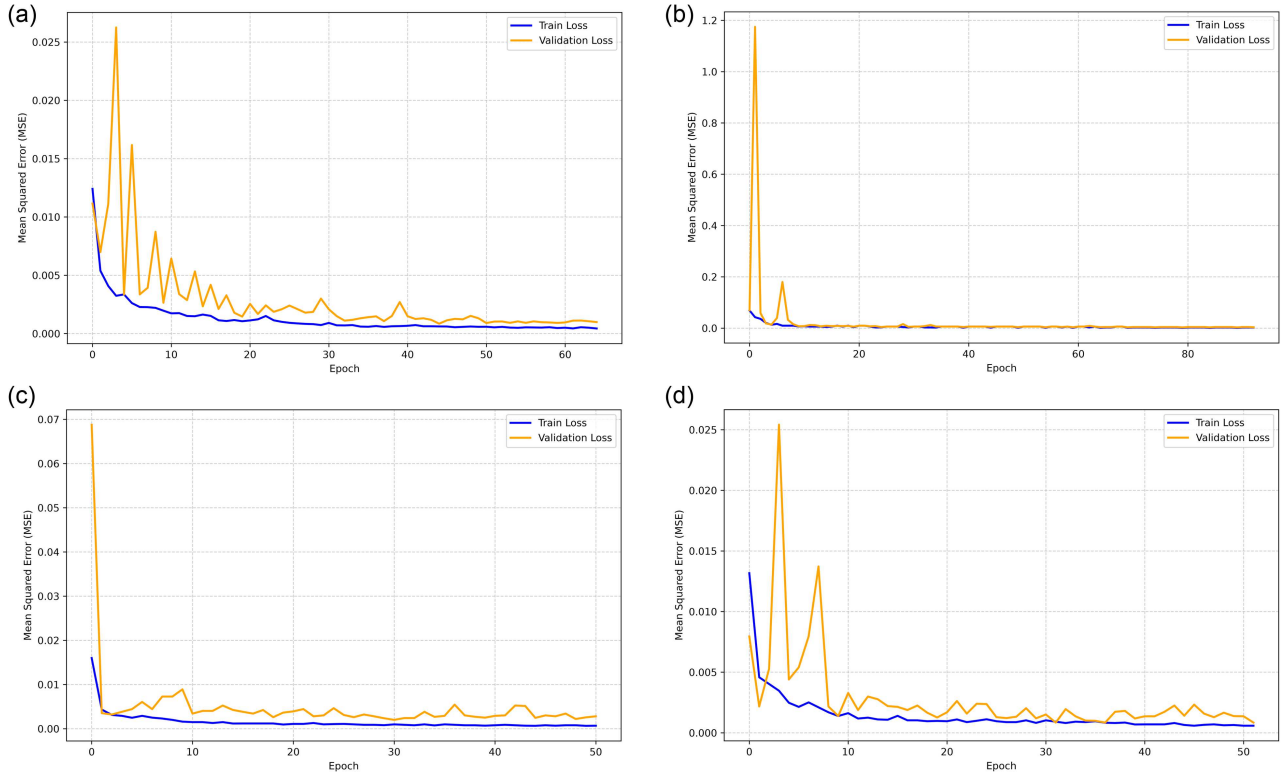
The overall study time also highlights the efficiency trade-off between the two. EfficientNet took the longest to train, taking 98 epochs. Since its "compound scaling" technique is more sophisticated, a lengthy setup was needed to get it optimal. On the other hand, these CNN (65 epochs), ResNet (50 epochs), and U-Net (50 epochs) converged in fewer epochs. This cooperative manner shows that all the frameworks tested adaptively learn from all features, and this dynamic process of learning, even when model complexity requires more training time than others, ensures that learning captures very complex nonlinear spatial and spectral relationships within the Jabodetabek landscape.

As summarized in Table 1, the quantitative evaluation across the independent test dataset indicates that all DL configurations achieve robust spatial generalization capabilities, with R^2 scores exceeding 0.94. Structurally, while the multi-scale spatial aggregations within the U-Net architecture secure the absolute peak accuracy metrics, the baseline performance of the CNN demonstrates that simpler, lightweight networks are highly viable and structurally efficient for mapping core physical urban distributions.

We show the quantitative evaluation of the models on an unseen test dataset in Table 1. Results All architectures show great generalization (all $R^2 > 0.62$), providing evidence that models can explain the vast majority of variation in urban density across the heterogeneous Jabodetabek region.

According to our latest results, the U-Net architecture achieves the best overall performance, securing the highest explanatory power ($R^2 = 0.9782967$) and the absolute lowest error metric ($MSE = 0.0006954$). The CNN came a close second as being competitive with an R^2 of 0.9748 and MSE of 0.0006998. Recent studies further support the notion that U-Net yields a high level of accurate segmentation, using the same architecture that was trained with 87% average accuracy on study

Figure 9
Training and validation curves: (a) CNN, (b) EfficientNet, (c) ResNet, and (d) U-Net



area segmentation of building footprint as city elements [28], validated multi-temporal metropolitan green space mapping with an average accuracy of 96.3% [29], and predicted impenetrable ground cover fractions globally achieving accompanied accuracies (92.46%) [30].

ResNet and EfficientNet (R^2 = 0.9683, R^2 = 0.9432) continued to perform well. Yet, the marginally lower performance of these models against the CNN and *U*-Net—as evidenced by their higher MSE and MAE scores—suggests that more complex, deeper core architectures do not always translate to a greater increase in accuracy for fractional regression problems. This offers an important counterpoint to recent trends; for example, [31] observed that transformer-based models (*e.g.*, Swin-U-Net) outperformed classical CNNs in capturing complex spatial details, but our findings show that a well-optimized U-Net or lean CNN can yield close-to-optimal results. This coincides with conclusions drawn by [32] that architectural depth cannot be the only way to achieve high performance where some urban land scene categories are concerned.

Crucially, the addition of classical ML baselines introduces a substantial benchmarking contrast. While both RF (R^2 = 0.6481569) and XGBoost (R^2 = 0.6263605) provide stable convergence, their descriptive capabilities remain notably inferior compared to any of the evaluated DL architectures. This distinct gap mathematically validates that deep representation learning is essential for high-fidelity mapping, as flattening spatial configurations into tabular mean metrics strips away local structural textures and pixel-to-pixel geometric hierarchies.

Additionally, as shown in Table 1, it can be observed that the predicted "Target Range" is different (0–0.849 for CNN and 0–0.898 for EfficientNet). This variation arises due to different saturation points of each model after inference on the same test set. All models obtained significant reductions in error compared with the baseline MSE, with ~40× reduction for both CNN and U-Net, confirming that DL indeed captures the complex nonlinear spatial and spectral relationships necessary to estimate urban density accurately.

The relationship between architectural complexity and predictive performance is presented in Table 2. One main finding was the relative failure of EfficientNet-B0, which achieved the largest RRMSE (39.74%), even compared with architectures that have relatively less advanced structural designs. The reason why this result was achieved lies in its "compound scaling" mechanism, which is mainly tuned for high-resolution natural image classification.

Given the task it is exposed to, urban density mapping, with input being a multispectral patch surrounding every pixel of interest, this complexity will probably lead to overfitting. It also suggests that the inefficient refinement of EfficientNet's complex feature-refinement mechanism may have introduced large amounts of model noise into its predictions, making it more difficult for this algorithm to extrapolate well across such a heterogeneous landscape as Jabodetabek compared with CNN, which extracts features in a much more straightforward manner given input data already enriched with spectral indices (NDVI and NDBI) explicitly representing built-up and vegetation characteristics.

The comparative footprints regarding trainable parameters, FLOPs, and operational runtime in Table 2 offer a compelling quantitative justification for adopting streamlined architectures. The most structurally complex deep architecture, U-Net (4.76 million parameters and 3.40 GFLOPs), yields the lowest RRMSE (24.59%), yet it provides only a marginal accuracy enhancement over the baseline performance of the CNN, which processes the fixed 128 × 128pixel spatial dimensions with far fewer constraints.

On the other hand, the CNN delivers a near-optimal predictive balance. It remains nearly four times lighter in structural payload and twice as fast during inference (0.4165 ms/sample) compared to its more complex U-Net and ResNet counterparts. Furthermore, the classical ML baselines (RF and XGBoost) demonstrate the extreme boundary of this computational trade-off; while they exhibit near-instantaneous training milestones (under 0.003 minutes) and negligible inference latencies, their predictive errors spike drastically ($RRMSE$>88%). For regional urban density mapping from multispectral imagery, these holistic benchmarks quantitatively demonstrate that escalating architectural complexity causes diminishing returns, wherein substantial surges in computational overhead yield only marginal, saturating improvements in localized predictive performance.

In Table 3, a paired t-test was executed on the absolute prediction errors to evaluate whether performance variations across models were statistically significant. The results indicate a distinct dichotomy in error distributions. The performance of the CNN is statistically unique and significantly different compared to all other architectures, confidently rejecting the null hypothesis ($p < 0.001$). Conversely, comparisons among the deeper networks, namely, EfficientNet vs. ResNet ($p = 0.126$), EfficientNet vs. U-Net ($p = 0.612$), and ResNet vs. U-Net ($p = 0.308$), fail to reject the null hypothesis, demonstrating that their error variances are not statistically significant.

These statistical bounds substantiate our conclusions regarding architectural diminishing returns. While U-Net achieves the absolute lowest error metrics, the high *p*-values among the heavier models confirm that expanding network depth or adding complex feature-refinement blocks yields functionally equivalent error behaviors. Crucially, the highly significant *p*-values associated with the CNN prove that its competitive performance is not due to stochastic sampling luck on the test dataset, validating it as a highly robust and computationally realistic option for large-scale urban density mapping.

Table 2
Quantitative comparison of model complexity, computational efficiency, and predictive accuracy

Metrics	CNN	EfficientNet	ResNet	U-Net	Random Forest	XGBoost
RRMSE	26.71%	39.74%	28.74%	24.59%	88.69%	91.40%
Params	1,208,353	4,009,981	4,932,673	4,757,313	N/A	N/A
Flops	0.8715 G	0.1403G	3.5197G	3.4039G	N/A	N/A
Inf. Time	0.4165 ms/sample	0.5511 ms/sample	0.9986 ms/sample	0.7842 ms/sample	0.0221 ms/sample	0.0015 ms/sample
Total Training Time	3.55 mins	7.27 mins	7.71 mins	5.38 mins	0.0026 mins	0.0008 mins

Table 3
Pairwise paired t-test matrix (p-values) computed from the models' absolute prediction errors

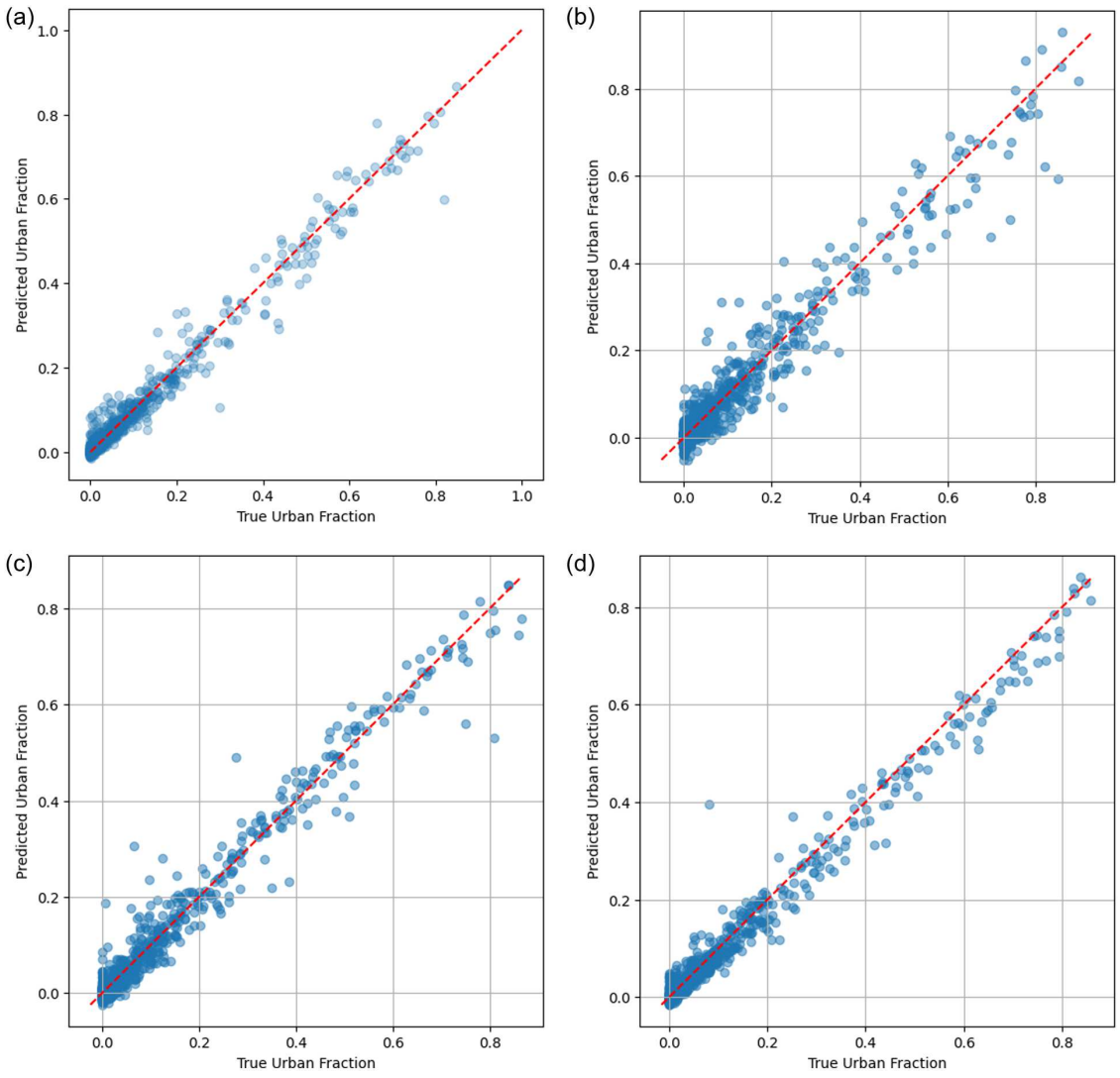
Model	CNN	EfficientNet	ResNet	U-Net
CNN	–	<0.001	<0.001	<0.001
Efficient Net	<0.001	–	0.126	0.612
ResNet	<0.001	0.126	–	0.308
U-Net	<0.001	0.612	0.308	–

The comparative data in Tables 2 and 3 shifts the argument for the simpler CNN from qualitatively based to quantitatively backed. Showcasing a statistically powered approach that results in a model with an order of magnitude fewer parameters and requiring 26 × fewer GFLOPs than previous versions achieves statistically better or equivalent results, paving the way for urban remotely sensed research to adopt sustainable computing practices. This reduces the computational burden and, consequently, the energy consumption and carbon footprint associated with training and deploying large ML models, thereby supporting UN Sustainable Development Goals 9 (Industry, Innovation and Infrastructure) and 13 (Climate Action).

Figure 10 includes a detailed view of regression performance for the four architectures studied. The distribution of data points over all subplots was strikingly close to the ideal diagonal line of 1:1 and exhibited significant prediction power over the varied urban landscape of Jabodetabek. The visualization of points accumulating in the lower range (0.0–0.4) corresponds to the distribution that predominates the test set, where all models could predict low-to-medium fractions of urban classes at a state with very high stability and little variation. The scatter decreases as the predicted values approach these limits.

In contrast, these subplots showed the different behavioral patterns associated with architectural complexity. Figure 10(a) and (d) present the most compact clustering around the diagonal, with the highest R^2 values of 0.9748 (CNN) and 0.9782 (U-Net), respectively. The linear trend of the fit indicates that these models may be used to predict the basic link between spectral signatures

Figure 10
Prediction vs. ground truth per model: (a) CNN, (b) EfficientNet, (c) ResNet, and (d) U-Net



and physical built-up density with little error and are very efficient. On the other hand, the plots of Figure 10(b) and (c) reveal more dispersion of residuals as the urban density increases. A larger value suggests a higher degree of heteroscedasticity, implying that the predictive variance increases in more complicated high-density metropolitan settings.

Figure 10 highlights the consolidated nature of recent designs; deeper models like EfficientNet and ResNet have a bigger capacity in theory but are also more quickly caught by local variance and were more affected by "model noise" for densities over ~900. On the other hand, U-Net and the simpler CNN were able to obtain a significantly more stable error profile over the whole range of urban fractions. Lastly, visually demonstrating that structural optimization, such as the U-Net's encoder–decoder transformation or direct extraction in CNN→regression, is typically the weakest link for consistency and correctness is more significant than architectural depth, even for successful multispectral-based urban density mapping.

Figure 11 (a)–(d) plot the residual error distributions to explore the statistical reliability of the four architectures. For all models, the residuals had no departures from zero in the center and a normal distribution approximating a stochastic error term, indicating that prediction errors were not related and that the models did not show any systematic bias. Such normalcy helps the models to successfully learn the underlying spatial and spectral fingerprints in Jabodetabek's heterogeneous region.

The subplots help in the comparison, as Figure 11(a) and (d) have much higher precision, both with taller and thinner residual peaks. This high concentration of errors near zero reflects their good predictive metrics, especially the R^2 of U-Net (0.9782) and CNN (0.9748). We see that both the encoder–decoder structure of U-Net and the more direct feature extraction of a CNN provide high-fidelity (low dispersion) regression.

However, the Figure 11(b) and (c) distributions were more dispersed, indicating larger MSE. The RRMSE values are presented in Tables 1 and 2. However, the increased dispersion of the residuals in both models implies that the increased architectural complexity might have contributed some "model noise" or over-parameterization, resulting in less consistent predictions than the performance achieved by the relatively basic CNN. It is notable that all models perform within an overall excellent range, but Figure 11 simply sets into consolidated visual evidence the conclusion that both U-Net and CNN are the strongest urban density estimation frameworks given their apparent reduction in error variance whilst maximizing predictive stability.

This allows for a detailed diagnosis of the model stability over the density spectrum. Figure 12 (a)–(d) show the distribution of absolute errors vs. the urban fraction for green and blue land covers. The absolute errors were mainly distributed in the low-to-medium urban fraction region (0.0 to 0.4) for all architectures, where excellent precision and consistency were found in the models. This is in accordance with the very low MAE values shown in Table 1, especially for Figure 12(d) and (a), with MAE values of 0.0156 and 0.0151, respectively.

Quantitative summaries of the subplots also showed varied levels of heteroscedasticity with variance in error increasing with target density. In panels (a) and (d), there is still some absolute error spread but not nearly as broad even at greater density ranges of the more advanced architectures, resulting in a considerably lower RRMSE value of 26.71% in the CNN architecture and 24.59% in U-Net, respectively. This stability suggests the capability of the models to make good use of spectral indices like NDVI

Figure 11
Residual error distribution per model: (a) CNN, (b) EfficientNet, (c) ResNet, and (d) U-Net

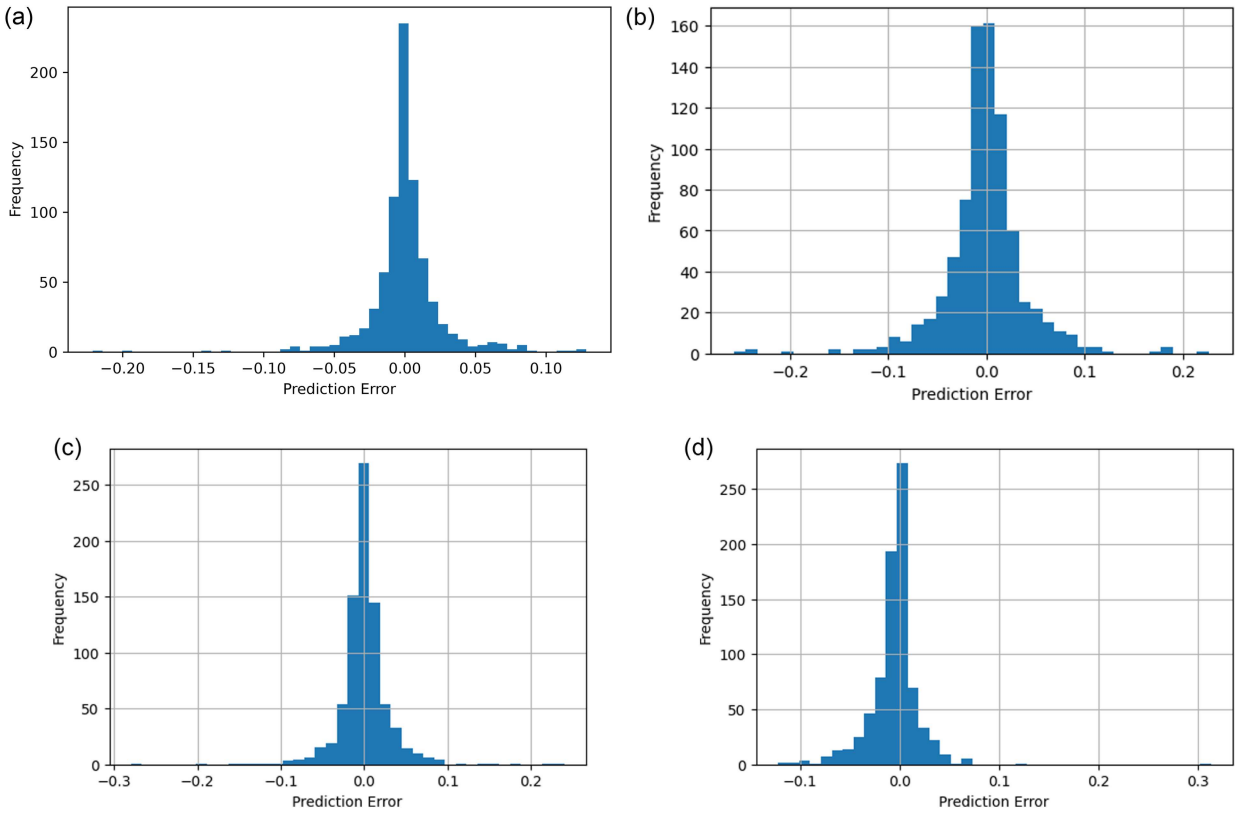
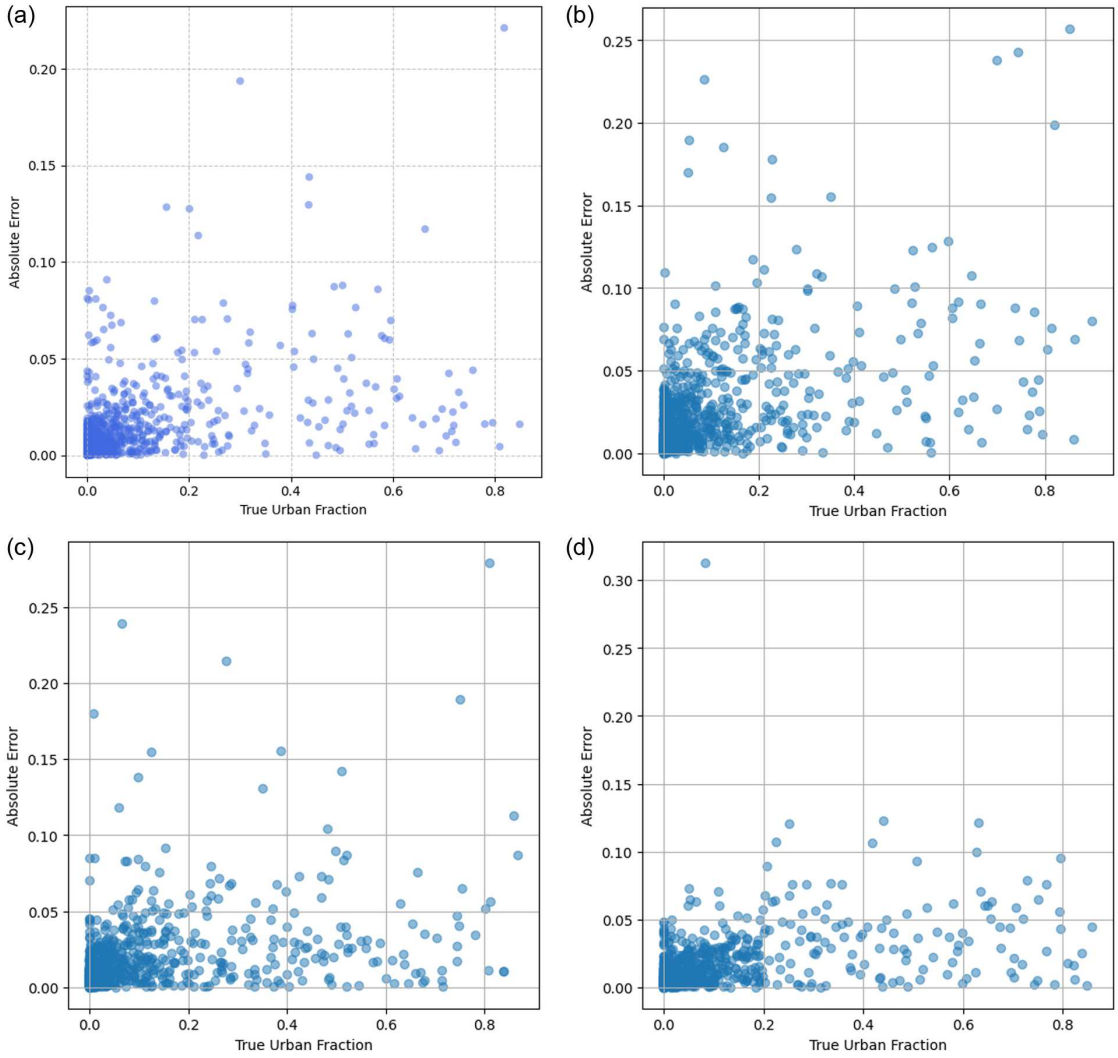


Figure 12
Absolute error distribution against ground truth per model: (a) CNN, (b) EfficientNet, (c) ResNet, and (d) U-Net



or NDBI to ensure reliable classification and less sensitivity to the structural complexity of dense urban blocks.

However, Figure 12(b) and (c) revealed considerably greater dispersion of absolute errors in mid-high ranges (above 0.5). The point spreads in these regions also appear to be bigger compared to the lower depth models, which is indicative of them entering a more variance-sensitive zone due to their greater architectural depth, aligning well with EfficientNet RRMSE at 39.74%. This structure of the distribution is consistent with the observation that high-capacity architectures lead to lower returns since deeper models provide additional "noise in the model," destabilizing local density estimations. Visually, from Figure 12, U-Net and the simpler CNN provide superior consistency in predicting quality vs. error over the varied Jabodetabek landscape.

Overall, this study provides empirical results that stimulate critical research on how to optimize architectural complexity under urban remote sensing challenges. Although the U-Net architecture provides the absolute highest accuracy overall, the CNN serves as a highly competitive alternative to traditional high-capacity regression frameworks, performing comparably while delivering considerably lower memory utilization and minimized latency per inference cycle. The structural coherence observed across the scatter plots, residual distributions, and absolute error patterns indicates that when multispectral data are heavily enriched with domain-specific spectral indices, resolving regional urban density fractions does not strictly necessitate excessive architectural depth. This directly demonstrates that a massive parameter count is not the primary requirement for effective urban monitoring at scale, but rather the structural viability of the network layout combined with robust, targeted feature engineering.

Nonetheless, we must rigorously acknowledge the geographical limitations of this regional study. Jabodetabek represents an organically growing, high-density Southeast Asian megacity framework. Consequently, our findings regarding model efficiency and spectral index utilization may not universally generalize to global urban morphologies with fundamentally contrasting spatial structures, such as rigid North American grid-based layouts or concentric European cores. While the absence of an independent external validation dataset across a different continent limits immediate cross-regional claims, this structural boundary explicitly frames spatial transpositionality and cross-regional transferability testing as a primary directive for our future research extensions. However, for spectrally and

geographically comparable fast-urbanizing regions, the streamlined CNN remains a highly viable trade-off between predictive skill and computational footprint boundaries.

5. Conclusion

We extensively evaluated four DL architectures, namely, CNN, EfficientNet, ResNet, and U-Net, to jointly regress urban density from multispectral Sentinel-2 data at a 10 m spatial resolution integrated with feature-engineered spectral indices. Quantitative assessment and visual diagnostics demonstrate that all DL configurations achieved stable convergence without structural overfitting, executing the spatial fractional regression task with high numerical fidelity. For each model, the distribution of prediction residuals was zero centered and contracted, demonstrating that the models are very effective in capturing complex nonlinear spatial and spectral interactions in heterogeneous Jabodetabek.

Empirical results showed that although the U-Net architecture achieved the maximum absolute accuracy (R^2 = 0.9782), the streamlined CNN emerges as a highly viable and high-performing alternative (R^2 = 0.9748958), delivering comparable predictive fidelity alongside a substantially minimized computational footprint. The investigation showed that the informative spectral indices (*e.g.*, NDVI and NDBI) satisfy most of the information required for the detection of built-up areas, and hence, there is no need for excessive architectural depth. Consequently, these empirical boundaries suggest a contextual trade-off regarding architectural complexity, demonstrating that expanding network depth or deploying highly intricate feature-refinement blocks yields diminishing returns and expanded prediction variances when processing spatial datasets already heavily enriched with targeted domain-specific inputs.

In conclusion, this work stresses that the choice of design should be a deliberate decision reflecting the trade-offs between predictive accuracy and computational and environmental economy. However, they are the most practical and scalable of their types, providing the high-fidelity results we had hoped for in a smaller footprint than the type of formula, though they cannot beat sophisticated algorithms such as object-based feature learning and 3D CNN in estimating impervious surfaces in the built environment under rapid development in urban regions because of their speed and weight at scale. While the quantitative outcomes presented in this work are specific to the complex, organically growing urban fabric of the Jabodetabek metropolitan region in Indonesia, they offer a strategic foundation for sustainable computing practices. Future investigations should rigorously evaluate the spatial transferability and transpositionality of these streamlined CNN designs across rigid, grid-based metropolitan layouts to comprehensively map architectural efficiency boundaries under completely distinct structural urban morphologies.

Recommendations

According to our study, the top two suggestions for planning future research on urban density assessment are (1) to explicitly address trade-offs between prediction power and computational efficiency of the model and (2) to enhance the ability of the model. The empirical evidence suggests that escalating network complexity does not uniformly translate to performance enhancements in spatial fractional regression. In this investigation, the streamlined CNN closely approached the peak predictive fidelity of the U-Net architecture, outperforming more complex models (EfficientNet and ResNet) with an R^2 of 0.9748958 while maintaining an optimized computational payload.

It also strongly suggests the extension of linguistic variables representing physical phenomena, derivative features such as the NDVI and the NDBI, which have a familiarity domain but do not belong to it. These indices are related spectral information that allows the model to more accurately predict built-up area fractions compared with raw satellite bands alone.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in the Google Earth Engine Data Catalog at https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR_HARMONIZED (Sentinel-2 Level-2A Harmonized), the Google Earth Engine Data Catalog at https://developers.google.com/earth-engine/datasets/catalog/ESA_WorldCover_v100 (ESA WorldCover v100), and Kaggle at https://www.kaggle.com/datasets/nwilhelm123/urban-density-fractional-regression (processed GeoTIFF files).

Author Contribution Statement

Willy Boen: Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Visualization. **Edy Irwansyah:** Conceptualization, Methodology, Formal analysis, Writing – review & editing, Supervision, Project administration.

References

[1] Yılmaz, M. (2023). A comparative assessment of the statistical methods based on urban population density estimation. *Geocarto International*, *38*(1), 2152494. https://doi.org/10.1080/10106049.2022.2152494

[2] Bao, W., Gong, A., Zhao, Y., Chen, S., Ba, W., & He, Y. (2022). High-precision population spatialization in metropolises based on ensemble learning: A case study of Beijing, China. *Remote Sensing*, *14*(15), 3654. https://doi.org/10.3390/rs14153654

[3] Neupane, B., Horanont, T., & Aryal, J. (2021). Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis. *Remote Sensing*, *13*(4), 808. https://doi.org/10.3390/rs13040808

[4] Huerta, R. E., Yépez, F. D., Lozano-García, D. F., Cobián, V. H. G., Fierro, A. L. F., de León, Gómez., H, Vargas-Martínez, A., … Vargas-Martínez, A. (2021). Mapping urban green spaces at the metropolitan level using very high resolution satellite imagery and deep learning techniques for semantic segmentation. *Remote Sensing*, *13*(11), 2031. https://doi.org/10.3390/rs13112031

[5] Arulananth, T. S., Kuppusamy, P. G., Ayyasamy, R. K., Alhashmi, S. M., Mahalakshmi, M., Vasanth, K., & Chinnasamy, P. (2024). Semantic segmentation of urban

environments: Leveraging U-Net deep learning model for cityscape image analysis. *PLoS ONE*, *19*(4), e0300767. https://doi.org/10.1371/journal.pone.0300767

[6] Seydi, S. T., & Arefi, H. (2023). A comparison of deep learning-based super-resolution frameworks for Sentinel-2 imagery in urban areas. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *X-1/W1–2023,* 1021–1026. https://doi.org/10.5194/isprs-annals-X-1-W1-2023-1021-2023

[7] 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), 3135. https://doi.org/10.3390/rs13163135

[8] Parekh, J. R., Poortinga, A., Bhandari, B., Mayer, T., Saah, D., & Chishtie, F. (2021). Automatic detection of impervious surfaces from remotely sensed data using deep learning. *Remote Sensing*, *13*(16), 3166. https://doi.org/10.3390/rs13163166

[9] Le, Q. H., Shin, H., Kwon, N., Ho, J., & Ahn, Y. (2022). Deep learning based urban building coverage ratio estimation focusing on rapid urbanization areas. *Applied Sciences*, *12*(22), 11428. https://doi.org/10.3390/app122211428

[10] Cai, J., & Chen, Y. (2022). A novel unsupervised deep learning method for the generalization of urban form. *Geo-Spatial Information Science*, *25*(4), 568–587. https://doi.org/10.1080/10095020.2022.2068384

[11] Fyleris, T., Kriščiūnas, A., Gružauskas, V., Čalnerytė, D., & Barauskas, R. (2022). Urban change detection from aerial images using convolutional neural networks and transfer learning. *ISPRS International Journal of Geo-Information*, *11*(4), 246. https://doi.org/10.3390/ijgi11040246

[12] Hao, M., Dong, X., Jiang, D., Yu, X., Ding, F., & Zhuo, J. (2024). Land-use classification based on high-resolution remote sensing imagery and deep learning models. *PLoS ONE*, *19*(4), e0300473. https://doi.org/10.1371/journal.pone.0300473

[13] Fayaz, M., Nam, J., Dang, L. M., Song, H.-K., & Moon, H. (2024). Land-cover classification using deep learning with high-resolution remote-sensing imagery. *Applied Sciences*, *14*(5), 1844. https://doi.org/10.3390/app14051844

[14] Jiang, F., & Ma, J. (2025). Predicting urban vitality at regional scales: A deep learning approach to modelling population density and pedestrian flows. *Smart Cities*, *8*(2), 58. https://doi.org/10.3390/smartcities8020058

[15] Liu, Q., Xu, J., Jiang, R., & Wong, W. H. (2021). Density estimation using deep generative neural networks. *Proceedings of the National Academy of Sciences*, *118*(15), e2101344118. https://doi.org/10.1073/pnas.2101344118

[16] Jin, B., & Xu, X. (2025). High-frequency CSI300 spot and futures price predictions via the neural network. *Journal of Uncertain Systems*, *18*(4), 2550008. https://doi.org/10.1142/S1752890925500084

[17] Suprijono, H., Kartikadarma, E., Yusianto, R., & Marimin, M. (2022). Defect detection of agricultural commodities using image processing and artificial neural networks. In *2022 International Seminar on Application for Technology of Information and Communication* (pp. 405–408). https://doi.org/10.1109/iSemantic55962.2022.9920445

[18] Jin, B., & Xu, X. (2024). Regional steel price index predictions for the southwest Chinese market through machine learning. *Ironmaking & Steelmaking: Processes, Products and Applications*, *0*(0), 03019233241297720. https://doi.org/10.1177/03019233241297720

[19] Jin, B., & Xu, X. (2025). A study of contemporaneous residential real estate price causation across major Jiangsu Province cities: Methodology using vector error-correction models and directed acyclic graphs. *Economics Open*, *1*, 2550008. https://doi.org/10.1142/s308284142550008x

[20] Madhiarasan, M., & Roy, P. P. (2022). *Hybrid transformer network for different horizons-based enriched wind speed forecasting*. arXiv. https://doi.org/10.48550/ARXIV.2204.09019

[21] Wang, F., Zhang, Y., Lei, C., Lu, Y., Zhao, Y., & Pan, W. (2024). Ensemble sparse approach to smart meter forecasting from different horizons. In *2024 5th International Conference on Computer Engineering and Application, 1762–1767*. https://doi.org/10.1109/iccea62105.2024.10604279

[22] More, N., Kanthe, A., Gangane, J., Padiya, P., & Sumithra, T. V. (2025). A complete study of remote sensing- Sentinel-2 satellite data for land use / land cover (LULC) analysis. *Panamerican Mathematical Journal*, *35*(1s), 231–249. https://doi.org/10.52783/pmj.v35.i1s.2311

[23] Wang, X., Zhang, Y., & Zhang, K. (2023). Automatic 10 m forest cover mapping in 2020 at China's Han river basin by fusing ESA Sentinel-1/Sentinel-2 land cover and Sentinel-2 near real-time forest cover possibility. *Forests*, *14*(6), 1133. https://doi.org/10.3390/f14061133

[24] Zhao, Q., & Qu, Y. (2024). The retrieval of ground NDVI (normalized difference vegetation index) data consistent with remote-sensing observations. *Remote Sensing*, *16*(7), 1212. https://doi.org/10.3390/rs16071212

[25] Muhaimin, M., Fitriani, D., Adyatma, S., & Arisanty, D. (2022). Mapping build-up area density using normalized difference built-up index (NDBI) and urban index (UI) wetland in the city Banjarmasin. *IOP Conference Series: Earth and Environmental Science*, *1089*(1), 012036. https://doi.org/10.1088/1755-1315/1089/1/012036

[26] Yasin, M. Y., Abdullah, J., Noor, N. M., Yusoff, M. M., & Noor, N. M. (2022). Landsat observation of urban growth and land use change using NDVI and NDBI analysis. *IOP Conference Series: Earth and Environmental Science*, *1067*(1), 012037. https://doi.org/10.1088/1755-1315/1067/1/012037

[27] Li, H., Rajbahadur, G. K., Lin, D., Bezemer, C.-P., & Jiang, Z. M. (2024). Keeping deep learning models in check: A history-based approach to mitigate overfitting. *IEEE Access*, *12*, 70676–70689. https://doi.org/10.1109/ACCESS.2024.3402543

[28] Irwansyah, E., Heryadi, Y., & Gunawan, A. A. S. (2020). Semantic image segmentation for building detection in urban area with aerial photograph image using U-Net models. In *2020 IEEE Asia-Pacific Conference on Geoscience, Electronics and Remote Sensing Technology*, 48–51. https://doi.org/10.1109/AGERS51788.2020.9452773

[29] Xu, L. L., Ma, K. S., Wang, X., Li, D. S., & Sun, H. (2024). Application of LA-UNet network model in remote sensing classification of urban green space. *Chinese Journal of Applied Ecology*, *35*(4), 1101–1111. https://doi.org/10.13287/j.1001-9332.202404.025

[30] Kuang, W., Hou, Y., Dou, Y., Lu, D., & Yang, S. (2021). Mapping global urban impervious surface and green space fractions using Google Earth Engine. *Remote Sensing*, *13*(20), 4187. https://doi.org/10.3390/rs13204187

[31] Sarhadi, A., Ravanshadnia, M., Monirabbasi, A., & Ghanbari, M. (2024). Optimizing concrete crack detection: An attention-based Swin U-Net approach. *IEEE Access*, *12*, 77575–77585. https://doi.org/10.1109/access.2024.3403389

[32] Sharma, M., Choudhary, R., & Anil, R. (2024). 2 T-UNet: A two-tower UNet with depth clues for robust stereo depth estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 757–764. https://doi.org/10.1109/cvprw63382.2024.00080

How to Cite: Boen, W., & Irwansyah, E. (2026). Beyond Complexity: Evaluating Simple vs. Complex Deep Learning Architectures for Urban Density Mapping. *Journal of Computational and Cognitive Engineering*. https://doi.org/10.47852/bonviewJCCE62029582