Digitizing Karachi's Decades-Old Cadastral Maps: Leveraging Unsupervised Machine Learning and GEOBIA for Digitization

Muhammad Waqas Ahmed
Urban & Infrastructure Engineering Department, NED University of Engineering & Technology, Karachi, Pakistan | UHasselt, Transportation Research Institute (IMOB), Martelarenlaan 42, 3500 Hasselt, Belgium
m.waqas.ahmed666@gmail.com (corresponding author)

Muhammad Ahmed
Urban & Infrastructure Engineering Department, NED University of Engineering & Technology, Karachi, Pakistan
muhammadahmed@neduet.edu.pk

Asif Ahmed Shaikh
Sukkur IBA University, Sukkur, Pakistan
asif.shaikh@neduet.edu.pk

Received: 18 March 2024 | Revised: 7 April 2024 | Accepted: 14 April 2024

 DOI: https://doi.org/10.48084/etasr.7280

ABSTRACT

In urban planning, land-use change is paramount for ensuring sustainable urban ecosystems. Monitoring, analyzing, and quantifying land use change is crucial to making statistical inferences and predicting the economic, environmental, and societal impacts of urban expansion. Recent technologies have enabled robust monitoring, recording, and documenting of spatio-temporal trends. When historical data remain nondigital, integrating modern technologies with traditional paper-based town maps becomes invaluable for digitization. Despite significant efforts in this field, little exploration has been done of the potential of Geographic Object-Based Image Analysis (GOBIA) for digitizing paper-based cadastral maps. This study introduces an innovative approach using unsupervised learning algorithms, K-means and Gaussian Mixture Models (GMM), in conjunction with GEOBIA techniques, to accurately extract land parcels from decades-old cadastral maps of Karachi, Pakistan. Initially, the maps were georeferenced using ArcGIS software, and unsupervised machine-learning algorithms were applied to preprocessed scanned images. Both clustering algorithms were evaluated based on key performance metrics, such as precision, recall, and F1 scores. The experimental results indicated that both algorithms performed well, with GMM slightly outperforming K-means in all aspects. GMM achieved 0.87 precision and recall and 0.86 F1 score of 0.86, while K-means achieved 0.82 precision, 0.78 recall, and 0.78 F1 score. Finally, unwanted features were removed by implementing a geometric criterion based on feature size and shape. This methodology effectively distinguishes between adjoining land parcels and ensures precise extraction of cadastral boundaries and land parcels, providing a reliable foundation for urban research and modeling.

Keywords-feature extraction; digital cadastre; historical maps; geographical information systems

I. INTRODUCTION

Urbanization is defined as the expansion of urban areas and urban-centric population growth. Previous studies have identified regional disparities in development mainly due to the focus of policymakers on urban development [1]. Recently, a study examined the urban expansion of 30 cities, uncovering a trend showing that cities are now expanding at a rate twice the population growth. This underscores the need for pragmatic solutions to improve the general livability and sustainability of an urban ecosystem [2]. Several studies have also attributed urban-centric growth to poor policy decisions that are loosely related to population growth, such as the case in the Tarai region of Nepal [3]. The expansion of cities globally has resulted in corresponding economic growth, making urbanization a critical engine fueling economic growth. In some cases, urbanization has improved the lifestyle of the citizens but also has resulted in several drawbacks, such as high

emissions, consumption, and extreme land-centered expansion [4]. Geographical Information Systems (GIS) have enabled urban planners, engineers, and researchers to formulate intelligent strategies by modeling urban systems and simulating growth patterns to reach informed conclusions and develop practical solutions [5]. For a precise quantification of the socioeconomic and environmental impacts of uncontrolled urban development, it is essential to establish a baseline by documenting the intended land use and identifying the deviation from the original plan [6]. Without satellite data, existing land use can be mapped by leveraging past land records and paper-based spatial maps [7]. Conventional image digitization methods are laborious and prone to human error, leading to mistrust in manual methods [8]. Recent advances in spatial sciences and computer vision have made it easier to identify features in images and extract meaningful information [9]. Recently, several studies have explored the prospects of Geographic Object-Based Image Analysis (GEOBIA) and Machine Learning (ML) algorithms to extract built-up features from satellite imagery. However, there is still a significant gap related to digitizing historical and paper-based cadastral maps [10]. The cumbersome digitization process can be facilitated by integrating ML and GEOBIA, resulting in improved classification accuracy and reduced cost and time. ML has been extensively utilized in the field of remote sensing. GIS platforms provide built-in ML algorithms that enable researchers to adopt the Iterative Self-Organizing Data Analysis Technique (ISODATA) and the Maximum Likelihood Classification (MLC) methods [11]. In [12], the ISODATA algorithm was used to convert morphologically transformed images into clusters for road feature extraction. Authors in [13] used the K-means clustering algorithm on high-resolution quick-bird satellite images to classify them into four classes with 88.89% accuracy. In [14], the prospects of spectral indices for inland water body detection were discussed, comparing the Otsu thresholding method with the K-means/ISODATA clustering algorithms. Previous works on cadastral map digitization adopted the image processing algorithms of segmentation and vectorization [15]. Segmentation identifies symbols and characters using a pattern recognition approach and is carried out by obtaining topological information to construct lines. In [16], a framework was proposed for cadastral boundary detection using a cellular automata-based algorithm called Moore Neighborhood Tracing (MNT) to extract lines by connecting nodes from paper-based maps with black pixels on a white background. In addition to land parcel digitization, there is also significant research on the applications of Optical Character Recognition (OCR) methods to extract letters and symbols from handwritten scripts [17] and geographical maps [18].

This study presents a novel approach to the automatic detection of cadastral boundaries from paper-based maps by combining unsupervised ML models, i.e., Gaussian Mixture Models (GMM) and K-means, with a GEOBIA-based contrast split algorithm. The novelty of the proposed workflow lies in its efficiency in differentiating land parcels and cadastral features, facilitating the comparison between existing built-up regions and planning details available in cadastral maps. The results of this method also provide valuable insights into the performance of GMM and K-means algorithms. The resultant method is a fully automated workflow that can be used to develop accurate baseline maps to facilitate policy decisions.

II. MATERIALS AND METHODS

A. Study Area

A well-planned urban settlement in the town of North Nazimabad, Karachi, was selected as the Region Of Interest (ROI). The town of North Nazimabad was established in 1953 by the Karachi Improvement Trust, later transformed into the Karachi Development Authority (KDA), as a housing scheme, locally known as Taimuriya. The KDA is a governmental authority responsible for Karachi's housing, infrastructural, and urban facility development. It was established by the Pakistani government in 1957 to regulate and promote urban growth in Karachi, and its functions include planning and developing various schemes in Karachi such as roads, water supply systems, sewer and stormwater drainage systems, and residential and commercial development [19, 20]. Figure 1 shows the geographical extent of the KDA scheme 2. The scheme comprises 21 blocks with 11,874 residential, 473 commercial, and 202 amenity plots [21]. The acquired cadastral map was produced at a scale of 1:3960 by the KDA. This map includes cadastral boundaries of different plot categories (size and function), road boundaries, and labels. The key objective of this study is to extract vectorized plot parcels in the form of polygons while removing redundant features (i.e., labels and annotations).

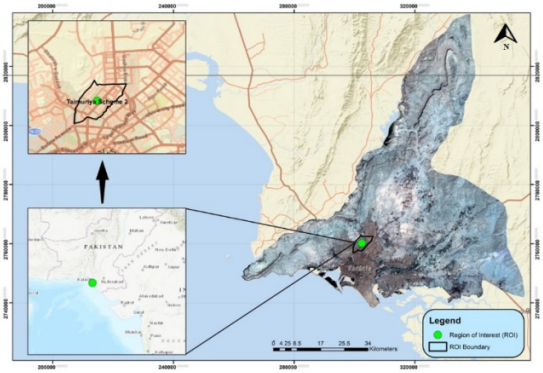

Fig. 1. The study's ROI (developed using the Arcmap software).

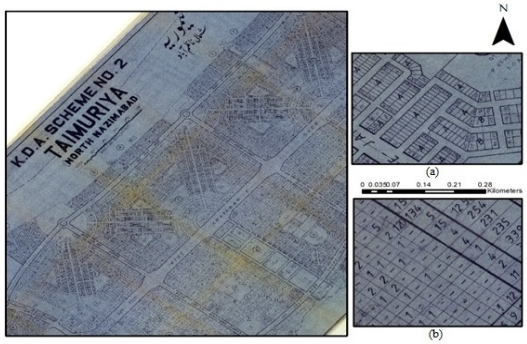

Fig. 2. Scanned cadastral maps of the ROI from the original master plan: (a) Plot boundaries and land parcels and (b) map labels and annotations posing challenges for efficient feature extraction.

B. Methodology

The proposed method involves a scanned map, manually georeferenced using GIS software to ensure accuracy and alignment with the current spatial reference. A Gaussian smoothing algorithm was applied to remove ink speckles and noise. Subsequently, unsupervised classification algorithms were applied to the preprocessed maps to identify and correctly extract cadastral boundaries and land parcels. The results were evaluated by comparing them with known accuracy reference maps, using precision, recall, F1 score, and Receiver Operating Characteristic (ROC) curve to gauge the accuracy of the resulting images. The proposed method ensures that the cadastral boundaries and land parcels extracted from the maps are accurate for baseline mapping.

1) Noise Removal by Gaussian Smoothing

Ink speckles and other noise were removed using the Gaussian smoothing function to avoid unwanted features during classification. Gaussian smoothing is used mainly for removing noise in digital signals by performing convolutions with a Gaussian function [20]. Its fundamental purpose is to replace each data point with the mean of neighboring data points weighted by a Gaussian function. The Gaussian function is used to describe phenomena such as probability distributions, signal processing, and image processing [21]. It is also known as the normal distribution and can be described as

$$f(x) = be^{-\frac{(x-c)^2}{2\sigma^2}} \quad (1)$$

where x is an integer representing the reflectance value in a digital number, b is the curve's height, c is the position of the peak center, e is Euler's constant (2.718), and σ is the standard deviation.

A Gaussian function is characterized by a bell-shaped symmetrical curve around its mean (μ). The highest point of the curve is at the mean. The spread of the distribution is determined by the standard deviation (σ), which is responsible for calculating the width. As the standard deviation increases, the peak becomes flatter [22]. Gaussian smoothing is widely applied in computer vision, digital image, and signal processing. Its main purpose is to eliminate noise, blur edges, and accentuate the relevant features present in the data. Gaussian smoothing is a critical preliminary step in image preprocessing, especially for object detection [23]. The prevalence of image noise can affect the overall classification accuracy and can result in a salt-and-pepper effect, due to high local spatial heterogeneity during spatial clustering.

2) Unsupervised Machine Learning

Unsupervised ML is a subset of ML that detects the underlying patterns within the data without using a labeled dataset. Unlike supervised learning, where labeled data are crucial for classification, an unsupervised learning algorithm aims to detect patterns, abnormalities, and relationships within the data and make the required inference accordingly [24]. Some of the most popular unsupervised ML algorithms include clustering, density estimation, anomaly detection, and dimensionality reduction [25]. Due to the nature of the input data, two popular unsupervised learning algorithms, K-means clustering and GMM, were used. Both algorithms yielded promising results, with GMM slightly outperforming K-means. Given the nature of the data and the computational expense of the ML algorithms, an image subset approach was adopted, as suggested in [26]. Classified images were stitched together after processing, making the processing more efficient by distributing the workload across the smaller subsets.

3) K-means Unsupervised Classification Algorithm

The K-means algorithm was applied to the filtered image, as the presence of noise can affect the overall accuracy of the classification. K-means [27] is a simple clustering algorithm. The algorithm initializes clusters based on the value of k, which is a user-defined hyperparameter [28]. The algorithm initially selects k points as centroids and subsequently assigns data points closest to these centroids based on their Euclidean distances. After the initial evaluation, the algorithm recalculates the centroids and adjusts the data points. This process was repeated until the centroids no longer moved [29]. Equation (2) shows the Euclidean distances between two points, p and q, in space, while the new centroid from the clustered groups is calculated using (3).

$$d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2} \quad (2)$$

$$centroid_i = \frac{1}{|S_i|}\sum_{x \in S_i} x_i \quad (3)$$

4) Gaussian Mixture Models (GMMs)

GMMs are parametric models defined as a weighted sum of Gaussian probability density functions [30]. In GMMs, data points are expected to follow a Gaussian or normal distribution with two key components: mean (μ) and standard deviation (σ). An Expectation-Maximization (EM) algorithm is used to fit the GMMs by identifying the distribution of each cluster. The EM algorithm iteratively estimates the maximum likelihood parameters in two steps: E and M. During the E step, the current parameters are used to estimate the posterior distribution of the variables, and based on this distribution, data points were assigned to each cluster. Subsequently, the model parameters were recalculated with a maximum likelihood rule [31].

A GMM learns the representation of multimodal distribution as a combination of unimodal distributions. Generally, GMMs work by parameterizing the cluster weights, mean, and covariance, where x is the cluster number [32]. If a dataset has K clusters, the GMM fits the data by the optimization equation:

$$\underset{a}{p \rightarrow} = \sum_{x=1}^{K} \phi_x N(a \mid \mu_x \sum_x) \quad (4)$$

where $\underset{a}{\rightarrow}$ models the probability density of the data point a as a mixture of K Gaussian (normal-N) distributions, each with its own mean μ_x, covariance, and mixing coefficient, ϕ_x are the weights for GMM's learning, and $\sum_x$ is the covariance matrix. Using the Gaussian distributions, datasets with limited noise levels can be refined. The applications of GMMs extend to various fields, such as pattern recognition, image segmentation, anomaly detection, speech recognition, and financial modeling [33].

C. Evaluation of Classification Performance

The confusion matrix is the most robust tool for evaluating the performance of ML classification models [34]. Confusion matrices evaluate the model's accuracy by comparing the predicted class with the true class labels [35]. The numbers of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) are used to measure the model's accuracy [36]. The parcels correctly classified are termed TP, while the parcels incorrectly classified as negatives are termed FN. Features classified correctly as negatives are TN and features incorrectly classified as positives are FP [37].

1) Recall

Recall is a criterion defined as detected positive samples out of all actual positive samples from a classification. A model with high recall suggests that it can effectively detect the positive samples from the class of interest [37]. Recall is calculated by:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (5)$$

2) Precision

Precision is the ratio between true positives and total positive predictions. An exact model ensures that it can accurately predict the positive samples. Precision is calculated by:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (6)$$

3) F-1 Score

The F1 score is a harmonic mean of precision and recall [38] and enables a balanced evaluation of the model's performance by considering both criteria, which can be calculated by:

$$\text{F1 score} = 2 * \frac{\text{Precision*Recall}}{\text{Precision+Recall}} \quad (7)$$

4) Receiver Operating Characteristics (ROC) Curve

A ROC curve graphically represents the performance of a classifier as its threshold discrimination varies. A ROC curve measures the true positive rate plotted against false positive rates and the Area Under the Curve (AUC). As shown in Table I, a value close to 1 is considered good, while an AUC of 0.5 is considered a random classifier [37].

TABLE I. AUC MEASURE SPECIFICATION ON CLASSIFICATION PERFORMANCE

Value	Performance
0.5 – 0.6	Poor
0.6 – 0.4	Fair
0.7 – 0.8	Good
0.8 – 0.9	Very Good
0.9 – 1.0	Excellent

D. Contrast Split Segmentation

Once the image was classified and evaluated, the resultant image was processed using a GEOBIA algorithm known as Contrast Split Segmentation (CSS) [39]. The CSS algorithm divides the image objects into distinct bright and dark areas [40]. An appropriate initial image object was determined using pixel filters with specific values for contrast and gradient to distinguish between objects. This algorithm was selected because of its ability to delineate objects based on variations in pixel intensity [41]. Once the image was classified into binary values of 0 and 1, the simplification dramatically facilitated the extraction of land parcels as objects and their subsequent geometric filtration.

III. RESULTS AND DISCUSSIONS

The proposed method was applied to 1:3960-scale cadastral maps produced by KDA, which are widely used by real estate agents and construction professionals. These maps delineate residential, commercial, and amenity plots, along with route maps for the residential scheme, accompanied by labels and tables showing plot counts for each residential block. Initially, the scanned RGB images were converted to grayscale through a mathematical conversion, resulting in grayscale images with pixel values ranging from 0 to 255. Values ranging from 0 to 145 signify the plot boundaries and written annotations, while 146 to 255 signify the land parcels within the boundaries. Gaussian filtering was applied to mitigate image noise [21].

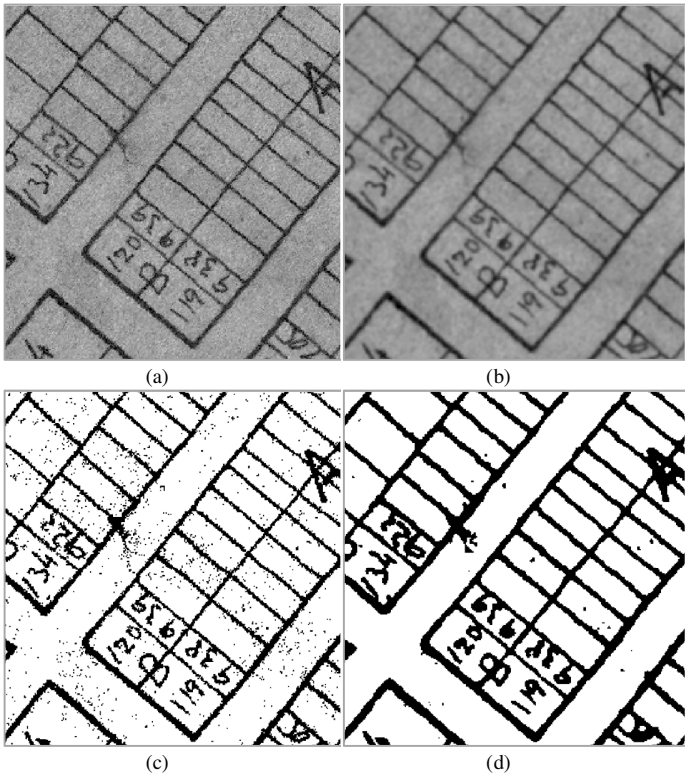

Fig. 3. The impact of noise removal on the overall classification of input images: (a) Grayscale image without Gaussian blur, (b) grayscale image with Gaussian blur, (c) classified image without Gaussian blur, (d) classified image with Gaussian blur.

Figure 4 and Table II demonstrate the superior performance of GMM over K-means. An important observation was the imbalanced class distribution throughout the experiment. This imbalance suggests that relying solely on the ROC metric may

not provide a comprehensive assessment, making the F1 score a more suitable evaluation metric. Table II presents the evaluation results for both classification algorithms on the specific dataset. The K-means algorithm achieved a precision of 82% and a recall of 78%, resulting in an F1 score of 78%. Similarly, the GMM algorithm achieved a precision of 87%, a recall of 87%, and an F1 score of 86%. The ROC-AUC metric for GMM was 79%, indicating its ability to distinguish between positive and negative samples.

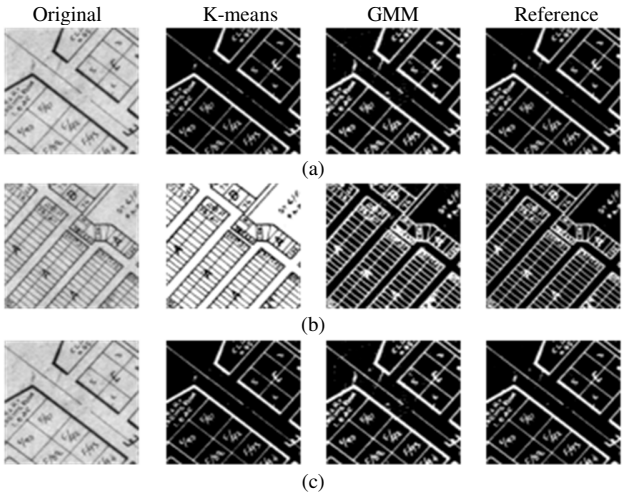

Fig. 4. Batch processing of the input image converted to smaller subsets. The left images show the original grayscale image after applying Gaussian blur followed by the two clustering algorithms (i.e., K-means and GMM).

TABLE II. CLASSIFICATION RESULTS

Algorithm	Precision	Recall	ROC-AUC	F1-Score
K-means	82%	78%	73%	78%
GMM	87%	87%	79%	86%

The results show that GMM outperformed K-Means in all metrics except ROC, which exhibited only a slight improvement. This suggests that GMM performed better than K-means in the given dataset [42]. However, due to the highly imbalanced class distribution, relying solely on the ROC metric may not provide a complete picture, making the F1 score a more appropriate evaluation metric [43, 44]. After obtaining the classified image with the most accurate results, achieved through the GMM, the vectorization process was carried out by segmenting the output image into distinct objects using the CSS algorithm. Several post-processing steps were performed using ArcGIS to enhance visual clarity and precision. These steps included label removal through vector merge operation, facilitating the seamless merging of overlapping polygons, and ensuring a clean representation of objects. Subsequently, the labels outside polygons were systematically eliminated based on geometric properties such as length, width, and area. This post-processing workflow significantly improved the quality and accuracy of the final segmented image, making it more suitable for subsequent analyses and applications.

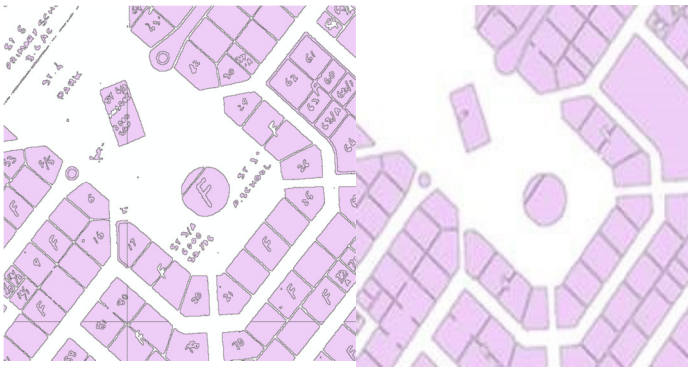

Fig. 5. Extracted features in a vector format: (a) The initial features with labels, (b) results from label removal through a feature merge operation. The labels inside the polygons were removed significantly.

IV. CONCLUSION

This study presented a method for extracting cadastral boundaries and land parcels from scanned maps of a region in Karachi. The proposed workflow used unsupervised classification algorithms to extract features from the preprocessed maps. K-Means and GMM were used, and their results were evaluated using classification metrics and reference maps. This study contributes to the existing body of knowledge by identifying the algorithm with better classification accuracy. In addition, the workflow identifies the most effective method for the removal of labels, annotations, and other unwanted features. In summary, the overall method demonstrated its efficiency in extracting individual land parcels compared to the approach that it directly drew inspiration from [16], where Moore's neighborhood tracing was utilized to identify intersecting points, requiring manual polyline drawing to establish cadastral boundaries. On the contrary, the proposed method significantly reduces the need for manual digitization tasks, thus minimizing any chance of human error in the vectorization process. The proposed method ensures the precision of the extracted cadastral boundaries and land parcels, establishing them as a reliable foundation for further research in urban modeling. This approach holds promise for applications in urban planning, land management, and infrastructure development, having the potential to facilitate real estate digitization and mapping by reducing associated costs and time.

REFERENCES

[1] L. Li and Y. Liu, "Spatial-Temporal Patterns and Driving Forces of Sustainable Urbanization in China Since 2000," *Journal of Urban Planning and Development*, vol. 145, no. 4, Dec. 2019, Art. no. 05019014, https://doi.org/10.1061/(ASCE)UP.1943-5444.0000528.

[2] S. Angel, J. Parent, D. L. Civco, A. Blei, and D. Potere, "The dimensions of global urban expansion: Estimates and projections for all countries, 2000–2050," *Progress in Planning*, vol. 75, no. 2, pp. 53–107, Feb. 2011, https://doi.org/10.1016/j.progress.2011.04.001.

[3] B. Rimal, L. Zhang, N. Stork, S. Sloan, and S. Rijal, "Urban Expansion Occurred at the Expense of Agricultural Lands in the Tarai Region of Nepal from 1989 to 2016," *Sustainability*, vol. 10, no. 5, May 2018, Art. no. 1341, https://doi.org/10.3390/su10051341.

[4] X. Guan, H. Wei, S. Lu, Q. Dai, and H. Su, "Assessment on the urbanization strategy in China: Achievements, challenges and

reflections," *Habitat International*, vol. 71, pp. 97–109, Jan. 2018, https://doi.org/10.1016/j.habitatint.2017.11.009.

[5] V. Maliene, V. Grigonis, V. Palevičius, and S. Griffiths, "Geographic information system: Old principles with new capabilities," *URBAN DESIGN International*, vol. 16, no. 1, pp. 1–6, Jan. 2011, https://doi.org/10.1057/udi.2010.25.

[6] T. W. Foresman, S. T. A. Pickett, and W. C. Zipperer, "Methods for spatial and temporal land use and land cover assessment for urban ecosystems and application in the greater Baltimore-Chesapeake region," *Urban Ecosystems*, vol. 1, no. 4, pp. 201–216, Dec. 1997, https://doi.org/10.1023/A:1018583729727.

[7] P. Drobež, M. K. Fras, M. Ferlan, and A. Lisec, "Transition from 2D to 3D real property cadastre: The case of the Slovenian cadastre," *Computers, Environment and Urban Systems*, vol. 62, pp. 125–135, Mar. 2017, https://doi.org/10.1016/j.compenvurbsys.2016.11.002.

[8] F. Döner, "Examination and comparison of mobile GIS technology for real time Geo-data acquisition in the field," *Survey Review*, vol. 40, no. 309, pp. 221–234, Jul. 2008, https://doi.org/10.1179/003962608X291013.

[9] R. Szeliski, Computer Vision: Algorithms and Applications. Springer Nature, 2022.

[10] B. Vaienti, R. Petitpierre, I. di Lenardo, and F. Kaplan, "Machine-Learning-Enhanced Procedural Modeling for 4D Historical Cities Reconstruction," *Remote Sensing*, vol. 15, no. 13, Jan. 2023, Art. no. 3352, https://doi.org/10.3390/rs15133352.

[11] S. Ul Din and H. W. L. Mak, "Retrieval of Land-Use/Land Cover Change (LUCC) Maps and Urban Expansion Dynamics of Hyderabad, Pakistan via Landsat Datasets and Support Vector Machine Framework," *Remote Sensing*, vol. 13, no. 16, Jan. 2021, Art. no. 3337, https://doi.org/10.3390/rs13163337.

[12] M. W. Ahmed, S. Saadi, and M. Ahmed, "Automated road extraction using reinforced road indices for Sentinel-2 data," *Array*, vol. 16, Dec. 2022, Art. no. 100257, https://doi.org/10.1016/j.array.2022.100257.

[13] B. Usman, "Satellite Imagery Land Cover Classification using K-Means Clustering Algorithm Computer Vision for Environmental Information Extraction," *Elixir Computer Science & Engineering*, vol. 63, pp. 18671–18675, 2013.

[14] H. Xie, X. Luo, X. Xu, H. Pan, and X. Tong, "Evaluation of Landsat 8 OLI imagery for unsupervised inland water extraction," *International Journal of Remote Sensing*, vol. 37, no. 8, pp. 1826–1844, Apr. 2016, https://doi.org/10.1080/01431161.2016.1168948.

[15] L. H. Lee and T. T. Su, "Vision-Based Image Processing of Digitized Cadastral Maps," *Photogrammetric Engineering & Remote Sensing*, vol. 62, no. 5, pp. 553–538, May 1996.

[16] A. Balkoca, A. İ. Yergök, and S. Yücekaya, "Vectorization of cadastral maps using image processing algorithms," in *2011 IEEE 19th Signal Processing and Communications Applications Conference (SIU)*, Antalya, Turkey, Apr. 2011, pp. 900–903, https://doi.org/10.1109/SIU.2011.5929797.

[17] M. G. Kibria and Al-Imtiaz, "BengalI Optical Character Recognition using self organizing map," in *2012 International Conference on Informatics, Electronics & Vision (ICIEV)*, Dhaka, Bangladesh, May 2012, pp. 764–769, https://doi.org/10.1109/ICIEV.2012.6317479.

[18] I. Schlegel, "Automated Extraction of Labels from Large-Scale Historical Maps," *AGILE: GIScience Series*, vol. 2, pp. 1–14, Jun. 2021, https://doi.org/10.5194/agile-giss-2-12-2021.

[19] A. Hasan, "Land contestation in Karachi and the impact on housing and urban development," *Environment and Urbanization*, vol. 27, no. 1, pp. 217–230, Apr. 2015, https://doi.org/10.1177/0956247814567263.

[20] A. Makandar and B. Halalli, "Image enhancement techniques using highpass and lowpass filters," *International Journal of Computer Applications*, vol. 109, no. 14, pp. 12–15, Jan. 2015.

[21] H. Kobayashi, B. L. Mark, and W. Turin, Probability, Random Processes, and Statistical Analysis: Applications to Communications, Signal Processing, Queueing Theory and Mathematical Finance. Cambridge University Press, 2011.

[22] M. J. Adams, *Chemometrics in Analytical Spectroscopy*. Royal Society of Chemistry, 2004.

[23] J. Šťastný and M. Minařík, "A Brief Introduction to Image Pre-Processing for Object Recognition," 2007.

[24] S. Rauschert, K. Raubenheimer, P. E. Melton, and R. C. Huang, "Machine learning and clinical epigenetics: a review of challenges for diagnosis and classification," *Clinical Epigenetics*, vol. 12, no. 1, Apr. 2020, Art. no. 51, https://doi.org/10.1186/s13148-020-00842-4.

[25] I. Syarif, A. Prugel-Bennett, and G. Wills, "Unsupervised Clustering Approach for Network Anomaly Detection," in *Networked Digital Technologies*, Dubai, United Arab Emirates, 2012, pp. 135–145, https://doi.org/10.1007/978-3-642-30507-8_13.

[26] F. Erdem and U. Avdan, "Comparison of Different U-Net Models for Building Extraction from High-Resolution Aerial Imagery," *International Journal of Environment and Geoinformatics*, vol. 7, no. 3, pp. 221–227, Dec. 2020, https://doi.org/10.30897/ijegeo.684951.

[27] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982, https://doi.org/10.1109/TIT.1982.1056489.

[28] X. Wang, P. Wu, Q. Xu, Z. Zeng, and Y. Xie, "Joint image clustering and feature selection with auto-adjoined learning for high-dimensional data," *Knowledge-Based Systems*, vol. 232, Nov. 2021, Art. no. 107443, https://doi.org/10.1016/j.knosys.2021.107443.

[29] C. K. Reddy and B. Vinzamuri, "A Survey of Partitional and Hierarchical Clustering Algorithms," in *Data Clustering*, Chapman and Hall/CRC, 2014.

[30] D. Reynolds, "Gaussian Mixture Models," in *Encyclopedia of Biometrics*, S. Z. Li and A. K. Jain, Eds. Boston, MA, USA: Springer US, 2015, pp. 827–832.

[31] M. R. Gupta and Y. Chen, "Theory and Use of the EM Algorithm," *Foundations and Trends® in Signal Processing*, vol. 4, no. 3, pp. 223–296, Apr. 2011, https://doi.org/10.1561/2000000034.

[32] S. Misra, H. Li, and J. He, *Machine Learning for Subsurface Characterization*. Gulf Professional Publishing, 2019.

[33] K. K. Verma, B. M. Singh, and A. Dixit, "A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system," *International Journal of Information Technology*, vol. 14, no. 1, pp. 397–410, Feb. 2022, https://doi.org/10.1007/s41870-019-00364-0.

[34] D. Gutierrez-Rojas, I. T. Christou, D. Dantas, A. Narayanan, P. H. J. Nardelli, and Y. Yang, "Performance evaluation of machine learning for fault selection in power transmission lines," *Knowledge and Information Systems*, vol. 64, no. 3, pp. 859–883, Mar. 2022, https://doi.org/10.1007/s10115-022-01657-w.

[35] D. Krstinić, M. Braović, L. Šerić, and D. Božić-Štulić, "Multi-label Classifier Performance Evaluation with Confusion Matrix," in *Computer Science & Information Technology*, Jun. 2020, pp. 1–14, https://doi.org/10.5121/csit.2020.100801.

[36] N. Wang, N. N. Zeng, and W. Zhu, "Sensitivity, Specificity, Accuracy, Associated Confidence Interval And ROC Analysis With Practical SAS Implementations," in *NESUG proceedings: health care and life sciences*, Baltimore, MD, USA, 2010.

[37] M. Bekkar and D. H. K. Djemaa, "Evaluation Measures for Models Assessment over Imbalanced Data Sets," *Journal of Information Engineering and Applications*, vol. 3, no. 10, pp. 27–29, 2013.

[38] D. Zhang, J. Wang, and X. Zhao, "Estimating the Uncertainty of Average F1 Scores," in *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, Northampton, MA, USA, Jun. 2015, pp. 317–320, https://doi.org/10.1145/2808194.2809488.

[39] M. K. Villareal and A. F. Tongco, "Remote Sensing Techniques for Classification and Mapping of Sugarcane Growth," *Engineering, Technology & Applied Science Research*, vol. 10, no. 4, pp. 6041–6046, Aug. 2020, https://doi.org/10.48084/etasr.3694.

[40] N. Mesner and K. Ostir, "Investigating the impact of spatial and spectral resolution of satellite images on segmentation quality," *Journal of Applied Remote Sensing*, vol. 8, no. 1, Jan. 2014, Art. no. 083696, https://doi.org/10.1117/1.JRS.8.083696.

[41] A. M. El-naggar, "Determination of optimum segmentation parameter values for extracting building from remote sensing images," *Alexandria*

Engineering Journal, vol. 57, no. 4, pp. 3089–3097, Dec. 2018, https://doi.org/10.1016/j.aej.2018.10.001.

[42] H. S. Kuyuk, E. Yildirim, E. Dogan, and G. Horasan, "Application of *k*-means and Gaussian mixture model for classification of seismic activities in Istanbul," *Nonlinear Processes in Geophysics*, vol. 19, no. 4, pp. 411–419, Aug. 2012, https://doi.org/10.5194/npg-19-411-2012.

[43] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing Imbalanced Data–Recommendations for the Use of Performance Metrics," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, Geneva, Switzerland, Sep. 2013, pp. 245–251, https://doi.org/10.1109/ACII.2013.47.

[44] D. Virmani, N. Jain, A. Srivastav, M. Mittal, and S. Mittal, "An Enhanced Binary Classifier Incorporating Weighted Scores," *Engineering, Technology & Applied Science Research*, vol. 8, no. 2, pp. 2853–2858, Apr. 2018, https://doi.org/10.48084/etasr.1962.