ICDSOS
The 3rd International Conference on Data Science and Official Statistics
November 27 - 28, 2025

"From Data to Impact: Leveraging Innovation in Data Science and Official Statistics for Data-Driven Policy Making and Global Competitiveness"

From Noisy Data to Insight: SOM Filtering Implementation For Improving the Machine Learning Model

Achmad Fauzi Bagus Firmansyah[1*]

[1] Directorate of Census and Survey Methodology Development, Statistics Indonesia-BPS

*Corresponding author's email: achmad.firmansyah@bps.go.id

Abstract. The filtering of representative training data from Big Data are critical steps in developing machine learning models, particularly for official statistics. This study demonstrates the application of Self-Organizing Map (SOM) filtering for enhancing training data quality in remote sensing-based classification of paddy phenological stages using satellite data. By clustering the data, SOM identifies and filters representative samples, which further removing noise and irrelevancy. Following the filtering, comparison is conducted between several purity threshold scheme and non-filtering dataset during model development. Findings reveal that increasing the purity threshold consistently improves classification performance and accuracy respectively, as filtering becomes stricter. The results demonstrate SOM filtering as an effective strategy for improving the representativeness and reliability of training datasets in remote sensing applications, while emphasizing the trade-offs when optimizing machine learning model robustness and generalizability.

Keyword: Representative Data Filtering, Self-Organizing Map (SOM)

1. Introduction

The filtering and selection of appropriate Big Data are critical steps in building high-quality training datasets for machine learning applications, particularly in the production of official statistics (UNECE, 2024). Typically, unfiltered Big Data contains irrelevant records, introducing noise and bias that can distort analyses and undermine the validity and performance of machine learning models. This challenge is especially pronounced for institutions such as National Statistical Offices (NSOs), since the use of suboptimal data may increase the risk of generating inaccurate statistical outputs (CBS, 2020). Nevertheless, despite these risks, practical experience and emerging use cases demonstrate that Big Data has become indispensable for capturing insights into modern phenomena—especially those that traditional data sources often fail to observe—resulting in more timely, granular, and cost-effective statistics. Thus, robust and systematic filtering procedures are not merely technical requirements but ethical imperatives in the data-driven era (ICO, 2017).

Numerous procedures exist for filtering Big Data to improve representativeness and quality in machine learning applications, ranging from basic steps such as outlier detection and duplicate removal to more advanced strategies such as prototype-based filtering (CORDIS, 2008; Wang et al., 2022; UNECE, 2024). Among these, Self-Organizing Map (SOM) filtering has emerged as a particularly prominent and effective approach for curating training datasets. By clustering high-dimensional data, SOM enables researchers to reliably identify and select representative records while filtering out noisy or irrelevant observations, thereby directly enhancing the quality of the resulting models (Li et al., 2021). This approach has been widely validated; for example, Bigdeli et al. (2022) applied SOM filtering to geochemical anomaly detection, and Cordel & Azcarraga (2015) used SOM to facilitate data visualization and filtering in large-scale datasets.

In remote sensing, SOM filtering is extensively applied to process high-dimensional satellite and environmental data, which are often affected by significant noise arising from atmospheric and sensor-related sources. If left unaddressed, such noise can obscure important features and reduce the reliability of machine learning model outputs. For instance, Hagen et al. (2003) employed SOM to cluster and filter satellite imagery, thereby identifying patterns and improving classification accuracy, while Guo et al. (2023) used SOM to assess spatiotemporal pollution patterns in remote sensing, enabling more representative sample selection and reducing dimensionality.

Building on the successful application of SOM filtering in previous research, this study implements SOM filtering within a Mixed Methods project conducted by Statistics Indonesia to classify paddy phenological stages from satellite imagery (BPS, 2025). By clearly delineating the SOM filtering procedure and its integration into the project workflow, this research demonstrates how the filtering process enhanced both the accuracy and overall performance of the classification model during its development, ultimately underscoring the valuable role of SOM in advancing remote sensing-based agricultural monitoring.

2. Research Method

The following section provides concept of the SOM algorithms, the corresponding data for the study, and the workflow of implementation of adopting SOM in the model development.

2.1. Concept of Self Organisation Maps (SOM)

A Self-Organizing Map (SOM), originally proposed by Kohonen (1982), is an unsupervised neural network that projects high-dimensional data into a low-dimensional discrete grid, preserving topological relationships. Each neuron on the grid possesses a weight vector equal in dimension to the data. During each training step, a data vector (***x***) is compared to all neuron weights, and the Best Matching Unit (BMU) is defined as the neuron with the smallest Euclidean distance to x:

$$c = arg\ arg\ | x \ - w_j| \qquad (1)$$

where $\boldsymbol{w_j}$ is the weight vector of neuron jj (Kohonen, 1982; Guérin et al., 2024; Hameed, 2024).

Once the BMU (***c***) is identified, the weights of both the BMU and its neighboring neurons are updated according to:

$$w_v(t+1) = w_v(t) + \alpha(t)\,\theta(u,v,t)\,\big(x(t) - w_v(t)\big) \qquad (2)$$

where $\alpha(t)$ is the learning rate, and $\theta(u,v,t)$ is the neighborhood function (often Gaussian or step function) centered at the BMU(u), controlling how strongly neighboring neurons are updated (Kohonen, 1982; Guérin et al., 2024; Hameed, 2024; van Heerden, 2024). As training progresses, the learning rate and the influence of the neighborhood decrease, guiding the map from global ordering to fine local adjustment.

For evaluating SOM implementation, two tools are commonly used: the U-matrix and quantization error. The U-matrix is a visual tool that displays the distances between neighboring neurons in the map, enabling users to easily identify cluster boundaries—well-formed clusters appear as distinct regions on the U-matrix, indicating strong SOM performance (Licen et al., 2023; Ultsch, 2003). In contrast, the quantization error provides a quantitative assessment of how closely the neurons' weights correspond to the original data points, where a lower value indicates that the SOM accurately represents its input data (Miftahuddin & Ridwan, 2025; Tu, 2015).

2.2. Data

The data used in this study consist of Satellite Imagery Time Series (SITS) of VV and VH polarizations from Sentinel-1A, collected over South Sumatra—the province with the largest rice harvest area on Sumatra Island (BPS Sumatera Selatan, 2023). During data acquisition, Sentinel-1A observations were filtered based on grids defined by the Area Sample Frame (ASF) method, a standard approach for agricultural monitoring in Indonesia (Amalia, 2019). The resulting dataset is biweekly and covers the period from late 2022 to 2024. For each grid segment, satellite data were integrated with monthly field observations; specifically, one month of field data was combined with the preceding four months of satellite observations to create a 10-value time series for each sample. To further enhance model performance, elevation data were included for each segment, although these values remained constant within each time series. This approach effectively leverages the strengths of SITS and ASF techniques, ensuring robust and objective agricultural monitoring in Indonesia (Gomarasca, 2019).

2.3. Workflow

The workflow of SOM Filtering implementation can be observed in following figure 1.

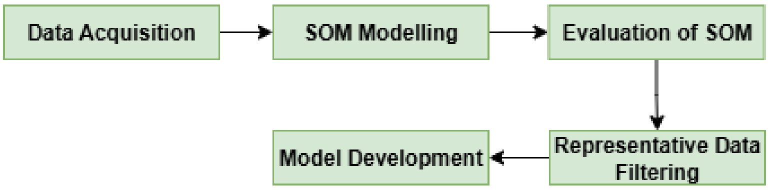

Figure 1. The workflow employed in this study.

This workflow begins with the collection of essential data, ensuring that all relevant information—such as satellite imagery and field survey results—is available for analysis. Next, the Self-Organizing Map (SOM) modelling stage processes the raw data to create organized clusters that reveal underlying patterns. Once the SOM is constructed, its performance is systematically evaluated to determine how effectively it captures and structures the data. Following this evaluation, the SOM is applied to filter and retain the most representative and reliable data points, ensuring that only high-quality information proceeds to the next stage. Finally, these carefully curated data samples serve as the foundation for subsequent model development, supporting accurate and meaningful analysis.

3. Result and Discussion

3.1. Pre-SOM Analysis

As a preliminary analysis, the distribution of data for each class was examined to identify the presence of imbalance. Figure 2 displays the data distribution based on city/regency and class. The heatmap reveals a heterogeneous distribution of record counts across the studied Kabupaten/Kota and phenological classes. Notably, Kabupaten/Kota 1607 stands out as a dominant contributor, particularly in the non-paddy class, which records the highest single count among all categories analyzed. Other vegetative classes, such as Tillage and Vegetative 1, also exhibit elevated counts in several regions, although these high values are concentrated in a limited subset of Kabupaten/Kota, underscoring region-specific agricultural activity. In contrast, many Kabupaten/Kota, especially those in the lower section of the matrix, consistently display low counts across all classes, suggesting either limited land use or reduced reporting activity. This spatial and categorical variability highlights a pattern of pronounced heterogeneity, indicating that a small number of regions account for the majority of activity in several vegetative classes, while many areas contribute minimally.

3.2. *SOM Analysis*

After distribution of the data is checked, the next step is applying SOM analysis for filtering the representative data. During the analysis, the following parameters is configured as configured in table 1.

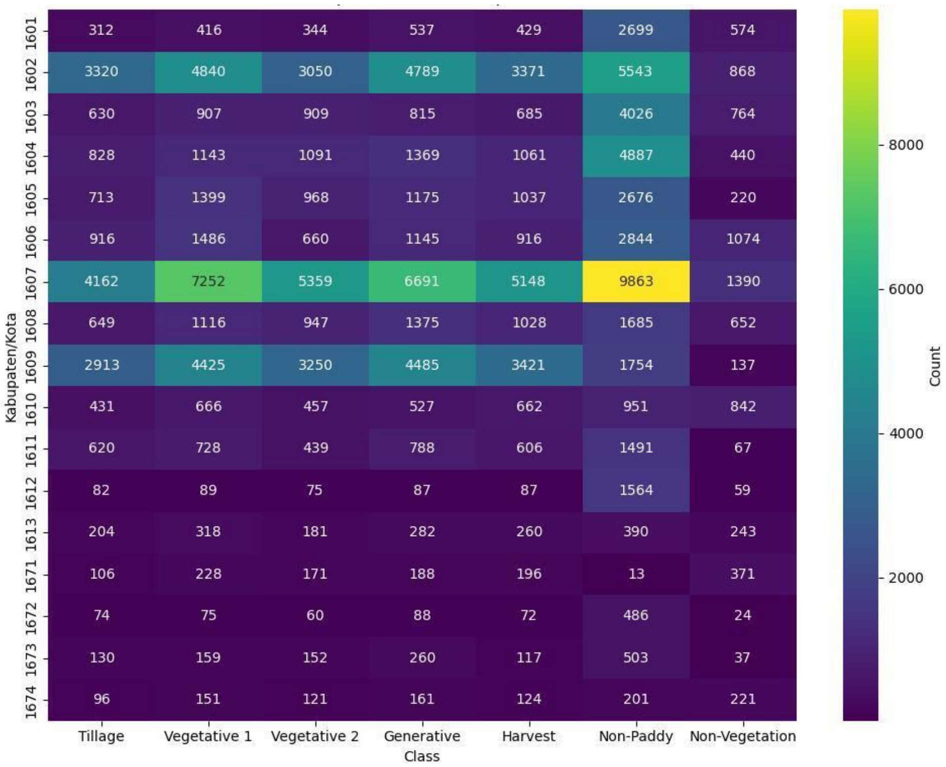

Figure 2. Heatmap of data distribution before SOM Analysis.

Table 1. Parameter for SOM.

Parameter	Value
Number of Neurons	10000 (with details 100 neurons in X and 100 neurons in Y)
Sigma	5
Learning Rate	0.5
Number of Iteration	5 times of the row counts

The selection of Self-Organizing Map (SOM) parameters is critical for optimal model performance. A large number of neurons (e.g., 10,000 arranged as 100 by 100) enables a finer representation of the input space and improved pattern identification, but it also increases computational demands and may result in unmapped neurons if the training data is sparse. The sigma parameter, which defines the neighborhood radius, must be carefully balanced—larger values promote smooth global topological preservation but can blur distinct clusters, while smaller values enhance cluster delineation at the risk of producing fragmented patterns. The learning rate, such as 0.5 in this study, governs the speed of adaptation; higher values accelerate convergence but may destabilize training if not properly decayed. The number of iterations, set here to five times the row count, aims to ensure thorough training, avoiding overly short training that yields incomplete maps or excessively long runs that increase computation without significant gains (Ahmed et al., 2019; Licen et al., 2023; Yin, 2008).

Subsequently, the evaluation was conducted by calculating the Quantization Error (QE) and plotting the U-Matrix. The QE, calculated as 2.8985, represents the average Euclidean distance between each data point and its closest neuron (best matching unit) on the SOM. This metric quantifies how well the map represents the given data: a lower QE indicates more accurate representation, while a higher QE suggests less precision. Further, the U-Matrix in figure 3 visualizes the mean distances between neighboring neurons on the trained SOM, where darker regions indicate similar neighboring neurons (potential clusters) and lighter areas represent greater distances, indicating possible cluster boundaries or transitions. The predominantly dark-to-moderate tones, interspersed with scattered lighter (yellow/green) spots, suggest that the data consists of numerous small or moderately defined clusters with many subtle transitions rather than large, sharply separated groups. This pattern indicates a complex or overlapping data structure, with contiguous dark regions representing areas of high similarity (clusters) and lighter regions marking the boundaries between different clusters or data types (Ultsch & Siemon, 1990; Licen et al., 2023).

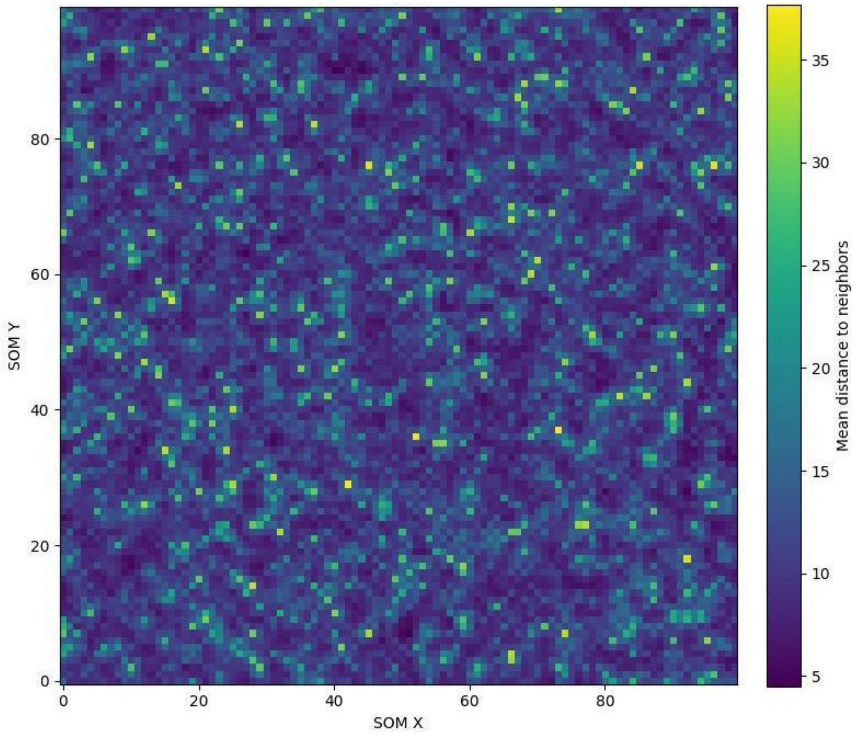

Figure 3. U-Matrix after SOM modelling conducted.

3.3. *Representative Analysis*

Following the development of the SOM model, each data point is mapped to its most similar neuron, known as the best matching unit (BMU) or winner. This assignment process is iterated until all data points are associated with their respective BMUs. Subsequently, for each neuron, the predominant class among the data points assigned to it is identified, enabling the classification of neurons based on the majority class they represent. This approach facilitates the interpretation of the SOM's structure and allows for the extraction of meaningful class labels from the clustered data.

Figure 4 illustrates the number of BMUs, which represent neurons or winners based on their dominant class. The horizontal bar chart displays the distribution of BMUs across the dominant classes on the SOM, revealing that Non-Paddy is the most prevalent and spatially extensive land cover type, while Non-Vegetation is the least represented, suggesting limited occurrence or coverage within the dataset. The intermediate BMU counts for Vegetative 1, Generative, Harvest, Tillage, and Vegetative 2 indicate a moderate spatial distribution of these classes and reflect overall landscape heterogeneity. Furthermore, the BMU class distribution in figure 5 demonstrates that all phenological classes are well-dispersed across the SOM grid in an interleaved pattern. This results in many small, mixed clusters rather than isolated class regions; in addition, the relatively uniform point size distribution suggests that the SOM effectively captures complex, overlapping data structures and context-dependent class transitions, consistent with observations in ecological and land classification SOM applications (Vesanto & Alhoniemi, 2000).

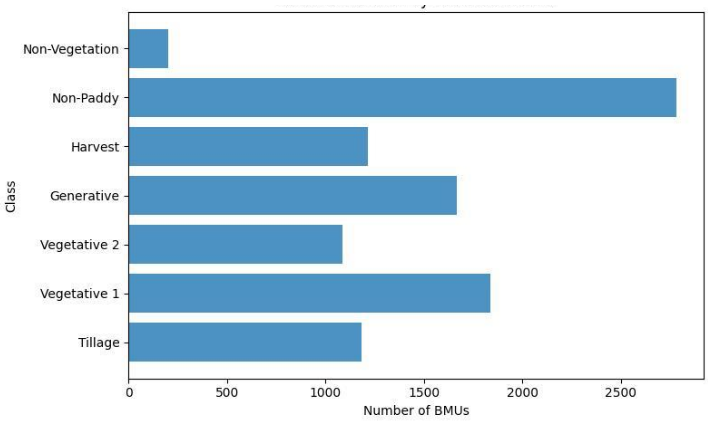

Figure 4. Number of BMUs (neurons) by dominant class.

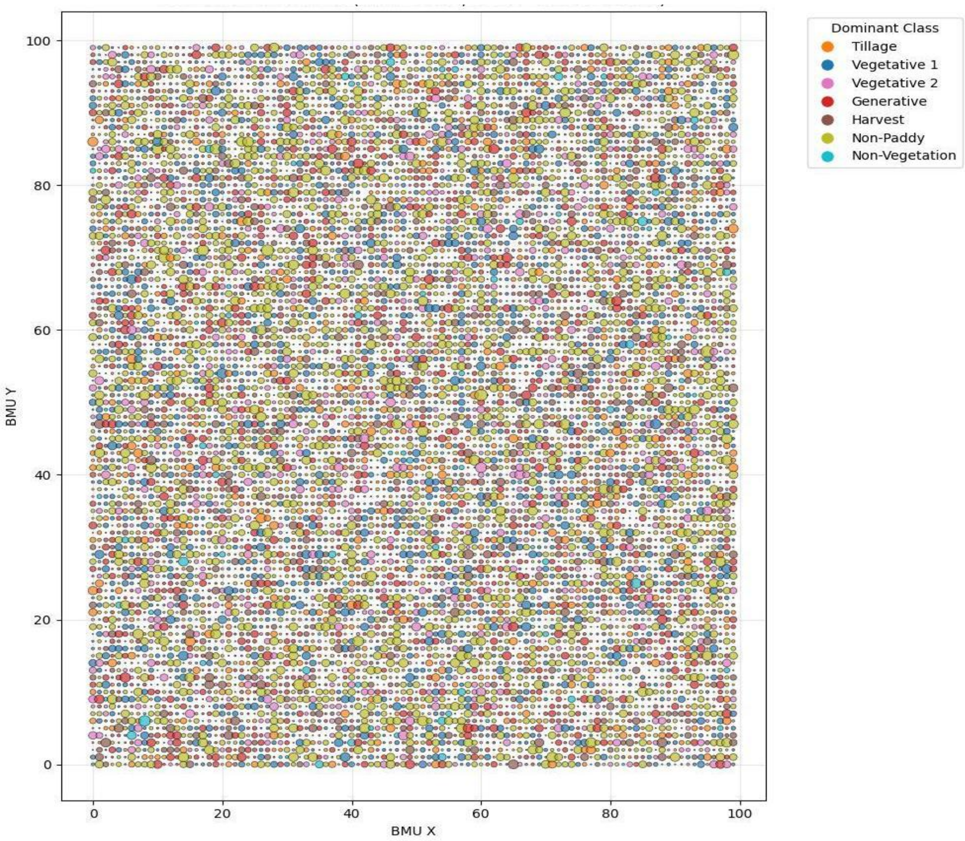

Figure 5. BMU distribution (size represent count proportionally color represent the dominant class).

3.4. Impurity Analysis and Filtering

The next step is impurity analysis for filtering the representative data as training data. The impurity analysis is employed by following procedures. Each BMU from the SOM is characterized by its dominant class and calculate the purity percentage, which identified by percentage of its majority class compared all class in one BMU. Reverse cumulative distribution function (CDF) plots are then produced for each class, measuring the number of rows affected by certain purity threshold.

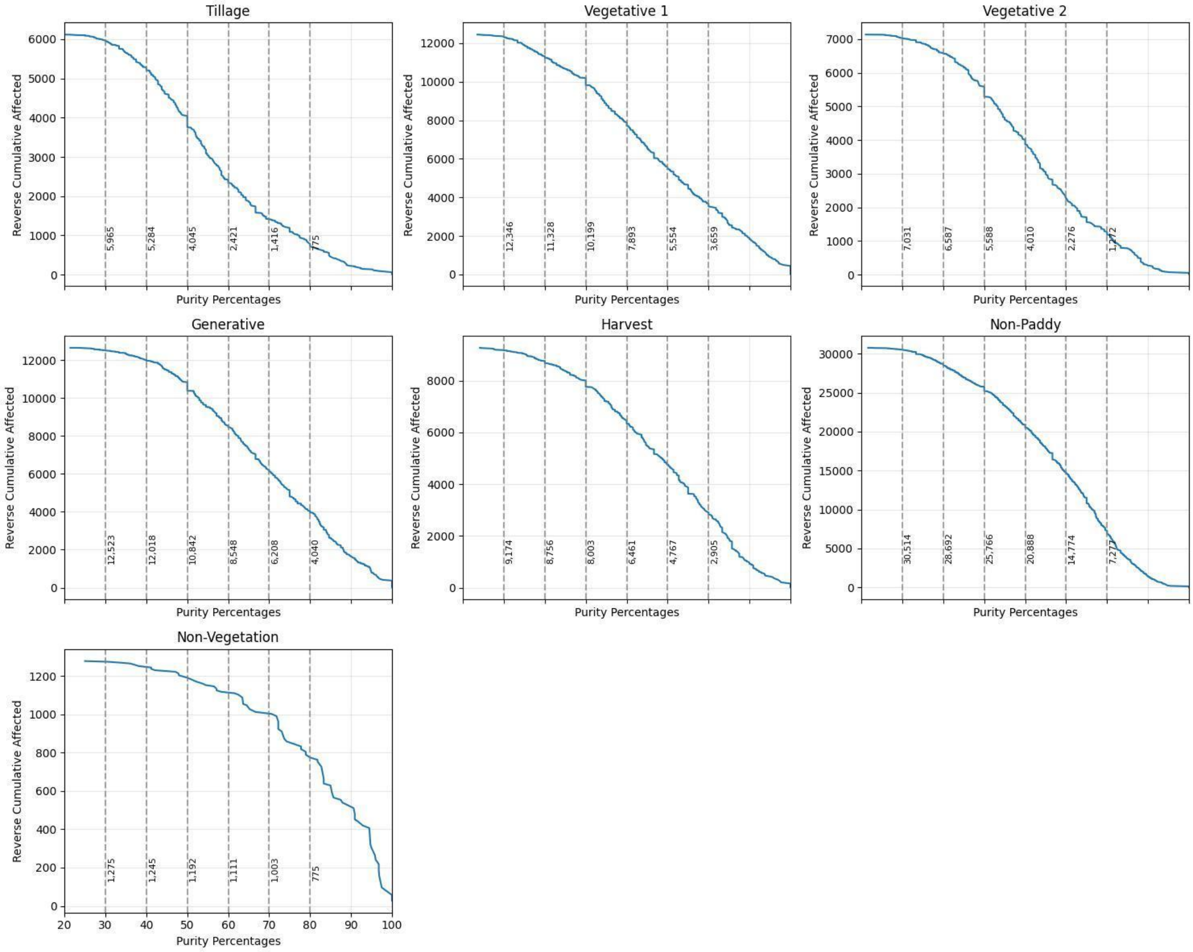

Figure 6. Purity percentages (reversed CDF) plots.

The reverse CDF subplots in figure 6 illustrate how the cumulative 'affected' metric decreases as the purity percentage increases for each dominant class. Each plot begins with the total class coverage at the lowest purity threshold (right side of the x-axis) and traces the decline in cumulative affected as the purity requirement becomes stricter (moving leftward toward higher purity). For all classes, including Non-Paddy and Vegetative 1, the curves exhibit a steady reduction, indicating that much of each class's representation derives from BMUs with moderate purity, while only a smaller fraction of affected data points is captured by the purest BMUs (those with the highest purity percentages). The gradual, extended decline rather than a sharp drop indicates that class coverage in this SOM is distributed across a continuum of purity levels, rather than being concentrated in a few exceptionally pure clusters. Classes such as Non-Vegetation, which have relatively small total affected values, approach zero more rapidly as purity increases, consistent with their lower presence and fewer high-purity nodes.

The heatmap in figure 7 displays the cumulative number of affected data points for each class as purity thresholds increase, with higher values indicating greater representation by highly pure BMUs. Non-Paddy consistently exhibits the largest affected totals across all purity levels, underscoring its dominance and strong, pure presence within the dataset. In contrast, Non-Vegetation and Tillage maintain the lowest cumulative affected values, reflecting limited and less pure representation in the SOM structure. As purity thresholds rise, all classes experience a marked decrease in cumulative affected, highlighting the challenge of preserving high representation at stricter purity levels. Based on these thresholds (30, 40, 50, 60, 70, and 80), the training data is filtered into several schemes, which are subsequently used in the modeling process.

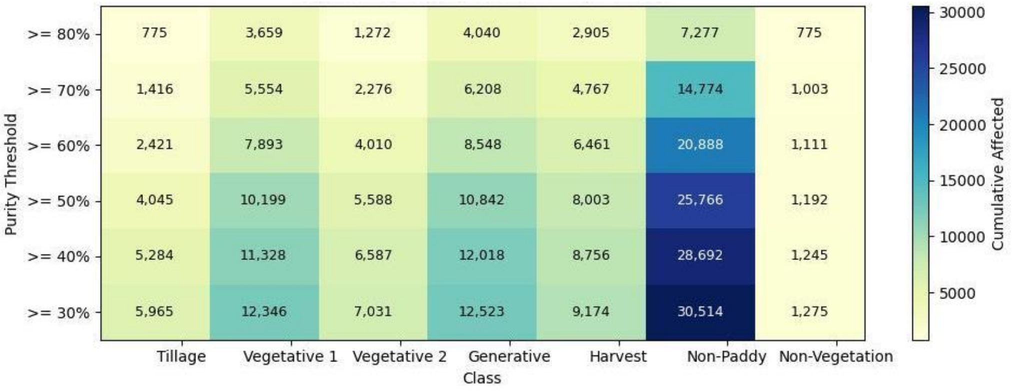

Figure 7. Data distribution after purity threshold applied.

3.5. *Impact Analysis to Modelling*

The impact of SOM-based filtering is evaluated by training classification models on datasets filtered using several dominant purity thresholds—specifically 30%, 50%, and 70%—as well as on the full, unfiltered dataset, which serves as a baseline. The XGBoost algorithm is selected for the modeling process due to its state-of-the-art performance on complex, structured data and its robustness in handling multiclass classification tasks (Chen & Guestrin, 2016). Feature selection for modeling aligns with that used in the SOM analysis to ensure consistency and comparability. For each threshold scenario, stratified 5-fold cross-validation is performed, utilizing a 70:30 split for training and validation while maintaining class proportion in each fold. To address class imbalance, class distribution-based weighting is applied, enhancing model fairness and predictive accuracy. Following cross-validation, key evaluation metrics—including accuracy, macro-averaged F1-score, macro precision, macro recall, and training time—are calculated to rigorously compare model performance across different levels of SOM filtering.

Table 2. Comparison results of SOM filtering in model development.

Purity Threshold	Number of Samples	Accuracy	F1-Macro	Precision-Macro	Recall-Macro	Runtimes (s)
0	79656	89.296%	87.138%	86.915%	87.388%	241.139
30	78828	89.542%	87.292%	87.030%	87.578%	244.859
50	65635	92.760%	90.520%	90.635%	90.422%	209.115
70	35998	97.228%	95.840%	96.131%	95.567%	151.051

The results, presented in table 2, clearly demonstrate that as the SOM purity threshold increases, the quality of classification improves significantly: validation accuracy rises from 0.89 (baseline, no filtering) to 0.97 at the highest threshold. Corresponding macro-averaged F1, precision, and recall metrics also exhibit marked gains. This improvement reflects that higher-purity clusters provide training samples with more distinct and less ambiguous feature-label relationships, enabling XGBoost to learn class boundaries more effectively and resulting in more reliable classification of all classes, including minority or phenologically similar categories.

A closer examination of the macro-averaged F1, precision, and recall reveals that each metric follows a similar upward trend, indicating broad improvements across all classes rather than improvements limited to the majority class. The macro-averaged F1-score increases from 0.87 to 0.96 as the purity threshold rises, reflecting a consistent balance between precision and recall regardless of class prevalence. Macro-averaged precision and recall exhibit nearly parallel growth, increasing from approximately 0.87–0.88 to over 0.96 at the 70% threshold, implying the model effectively minimizes false positives and captures nearly all actual positives, even in less common classes.

However, these gains come with a substantial reduction in available training data—from approximately 80,000 samples to just under 36,000 at the 70% threshold. While this reduction shortens training times, it may also increase the risk of overfitting, particularly when applied to even smaller or more heterogeneous datasets. At intermediate thresholds (e.g., 50%), there is a significant increase in both data quantity and quality, yielding notable improvements in accuracy and F1-score without an excessive loss of samples. This suggests moderate filtering can achieve an optimal balance between model robustness and generalizability (Zhou et al., 2020). These findings highlight the nuanced trade-off in supervised learning between maximizing clean, high-information training data and retaining sufficient sample diversity for effective generalization—an effect well documented in ensemble learning and class imbalance research (Chen & Guestrin, 2016; Buda, Maki, & Mazurowski, 2018).

3.6. Discussion and Reflection

The results of this study demonstrate a clear benefit of applying SOM-based filtering prior to model training, as evidenced by the consistent improvement in performance metrics with increasing purity thresholds. Retaining only the most representative and unambiguous samples enables XGBoost to more effectively capture the underlying class structure, thereby reducing misclassification and class ambiguity (Chen & Guestrin, 2016; Zhou, Sun, & Wang, 2020). Notably, enhancements in macro-averaged F1, precision, and recall at higher thresholds indicate more reliable predictions not only for the majority class but also for minority and transitional phenological classes. Moderate filtering thresholds (e.g., 50%) offer a practical compromise by preserving sufficient sample diversity to support generalization while simultaneously improving accuracy—a trade-off commonly discussed in the literature (Buda, Maki, & Mazurowski, 2018).

However, filtering out lower-purity samples may also exclude important edge cases, rare transitions, or ambiguous instances, potentially resulting in models that perform well on clean data but lack robustness in noisy, real-world scenarios. Future research could address these limitations by exploring hybrid approaches, such as ensemble modeling across different purity thresholds or integrating uncertainty quantification, alongside validating results on more diverse or independent datasets. Practically, these findings emphasize the critical role of domain expertise and careful threshold selection, as the optimal filtering strategy will ultimately depend on the specific objectives and requirements of each application.

4. Conclusion

This study has demonstrated the effective implementation of SOM filtering for curating high-quality training data in the remote sensing-based classification of paddy field phenological stages. Visualization of the Best Matching Unit (BMU) distribution revealed that the SOM successfully identified dominant land cover types, with Non-Paddy areas forming the most extensive and well-represented clusters, while Non-Vegetation was least represented. The BMU class distribution and purity plots illustrated a well-dispersed presence of all phenological classes across the SOM grid, highlighting the network's capability to discern both dominant and subtle patterns in the data. Applying different SOM filtering thresholds resulted in notable improvements in model accuracy and macro F1-score, increasing from 89% to 97% and 87% to 96%, respectively, as the purity threshold was raised. These gains were consistent across macro-averaged precision and recall metrics and were particularly effective in clarifying class boundaries and reducing misclassification, benefiting not only majority classes but also minority and transitional phenological classes. Overall, the model comparison confirmed that moderate filtering levels could substantially enhance both accuracy and data representation without excessively reducing the training sample size.

The application of stricter purity thresholds significantly improved data quality and classification performance but also led to a considerable reduction in available training samples, underscoring the inherent trade-off between data purity and class diversity. This reduction poses risks of overfitting and the exclusion of rare or transitional cases essential for robust, real-world applications. Therefore, future research should investigate hybrid data filtering methods, ensemble modeling strategies, or adaptive thresholding techniques to sustain high model reliability while preserving valuable data diversity. Further validation on more varied datasets and incorporation of uncertainty quantification are recommended to enhance model robustness and practical utility in large-scale agricultural monitoring systems.

References

[1] Ahmed, M., Salama, A. S. A., Rady, S., & Eissa, S. (2019). "Optimizing self-organizing maps parameters using adaptive differential evolution algorithm" *In 2019 15th International Computer Engineering Conference (ICENCO)* (pp. 70–75). IEEE.

[2] Amalia, R. R. (2019). "Developing crop cutting survey using area sampling frame" *MPRA Paper* No. 119487.

[3] Badan Pusat Statistik Sumatera Selatan. (2023). "Harvest area and rice production in Sumatra Selatan 2022." https://sumsel.bps.go.id (accessed in 25 August 2025)

[4] Badan Pusat Statistik (2025). *Laporan Mixed Method Tahun 2024*

[5] Bigdeli, A., Maghsoudi, A., & Ghezelbash, R. (2022). "Application of self-organizing map (SOM) and K-means clustering algorithms for portraying geochemical anomaly patterns."

[6] Buda, M., Maki, A., & Mazurowski, M. A. (2018). "A systematic study of the class imbalance problem in convolutional neural networks." *Neural Networks*, 106, 249–259.

[7] CBS. (2020). "Big data in official statistics". https://www.cbs.nl (accessed in 18 August 2025)

[8] Chen, T., & Guestrin, C. (2016). "XGBoost: A scalable tree boosting system." *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

[9] CORDIS. (2008). *D3.1 data filtering methods* - CORDIS.

[10] Cordel, M. O., & Azcarraga, A. P. (2015). "Fast emulation of self-organizing maps for large datasets."

[11] Gomarasca, M. A. (2019). "Sentinel for applications in agriculture." *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* XLII-3/W6, 91–98.

[12] Guo, L., et al. (2023). "Self-organizing map algorithm for assessing spatial and temporal pollution patterns in environmental remote sensing."

[13] Hagen, A., et al. (2003). "Using self-organizing maps to identify patterns in satellite imagery."

[14] ICO. (2017). "Big data, artificial intelligence, machine learning and data protection." https://ico.org.uk/ (accessed in 17 August 2025)

[15] Le Anh Tu. (2015). "Improving the quality of self-organizing map by "different elements" competitive strategy." *Journal of Computer Science and Cybernetics,* 31(3), 227–238.

[16] Li, Y., Zhang, X., & Shao, Z. (2021). "A hybrid visual tracking algorithm based on SOM."

[17] Licen, S., Carminati, E., Crocetti, E., & Pellizzari, M. (2023). " Self-organizing map algorithm for assessing spatial and temporal patterns in environmental data." *Science of The Total Environment*, 905, 167592.

[18] Miftahuddin, Y., & Ridwan, A. R. S. (2025). "Application of self-organizing map and K-means to cluster bandwidth usage patterns in campus environment." *JOIN: Jurnal Online Informatika,* 10(1), 66–76.

[19] Ultsch, A. (2003). "Maps for the visualization of high-dimensional data spaces." *In Proc. WSOM* (pp. 225–230).

[20] Ultsch, A., & Siemon, H. P. (1990). "Kohonen's self-organizing feature maps for exploratory data analysis". *In Proc. INNC'90, International Neural Network Conference* (pp. 305–308).

[21] UNECE. (2024). "The use of machine learning in official statistics." https://unece.org (accessed in 18 August 2025)

[22] Vesanto, J., & Alhoniemi, E. (2000). "Clustering of the self-organizing map." *IEEE Transactions on Neural Networks*, 11(3), 586–600

[23] Wang, H., Zhao, S., Chen, J., et al. (2022). "Prototype-representations for training data filtering in classification of big data."

[24] Yin, H. (2008). “The self-organizing maps: Background, theories, extensions and applications”. *In Computational Intelligence: Research Frontiers* (pp. 715–720). Springer. https://doi.org/10.1007/978-3-540-68830-3_84

[25] Zhou, Z., Sun, L., & Wang, J. (2020). “The impact of data quality on the performance of machine learning models”. *Data Science Journal*, 19(1), 6. https://doi.org/10.5334/dsj-2020-006