



An Improved Spectral Residual Algorithm for Sparrow Pest Detection in Rice Field CCTV Video

Raja Ayu Mahessya¹, Yuhandri², Rini Sovia³

¹Department of Information Engineering, Faculty of Computer Science, Universitas Putra Indonesia YPTK Padang, Padang, Indonesia

^{2,3}Department of Information Technology, Faculty of Computer Science, Universitas Putra Indonesia YPTK Padang, Padang, Indonesia

*ayumahessya@upiypk.ac.id

Abstract

Agricultural pest surveillance in open field environments presents a persistent challenge for food security, particularly when early detection is limited by the inadequacy of traditional monitoring methods. Existing Spectral Residual (SR) algorithms, while capable of saliency-based object detection, exhibit notable limitations in adapting to dynamic agricultural conditions resulting in degraded segmentation quality and elevated false-positive rates. This study proposes an improved Spectral Residual algorithm that independently processes Red, Green, and Blue (RGB) channels via Fast Fourier Transform (FFT), preserving richer spectral information that is otherwise lost in conventional grayscale-based approaches. The proposed pipeline integrates spatial saliency maps with temporal motion cues extracted through frame-difference analysis, suppressing static background noise while reinforcing responses in regions exhibiting genuine movement. Adaptive binarization and Support Vector Machine (SVM) classification are further incorporated to strengthen object segmentation robustness across diverse lighting and background conditions. The model was evaluated on 136 frames extracted from 2MP Hikvision CCTV footage installed at rice field sites, benchmarked against the classical Spectral Residual baseline using accuracy, precision, recall, and F1-score. The proposed method achieved an accuracy of 98.15%, precision of 91.67%, recall of 98.98%, and F1-score of 95.18% — representing substantial improvements over the baseline (74.00%, 70.00%, 65.00%, and 69.00%, respectively). These results demonstrate that multi-channel spectral processing combined with motion-aware saliency yields a robust and computationally efficient saliency framework whose low algorithmic complexity makes it a strong candidate for real-time sparrow pest detection in CCTV-based agricultural monitoring systems.

Keywords: accuracy; CCTV Video; digital image processing; pest detection; spectral residual algorithm

How to Cite: R. Ayu Mahessya, Y. Yuhandri, and R. Sovia, "An Improved Spectral Residual Algorithm for Sparrow Pest Detection in Rice Field CCTV Video", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 4, pp. 1164–1173, Aug. 2026, doi: [10.29207/resti.v10i7775](https://doi.org/10.29207/resti.v10i7775).
Permalink/DOI: <https://doi.org/10.29207/resti.v10i7775>

Received: April 30, 2026

Accepted: July 7, 2026

Available Online: August 23, 2026

This is an open-access article under the [CC BY 4.0 License](#)

Published by [Ikatan Ahli Informatika Indonesia](#)

1. Introduction

Digital image processing has fundamentally transformed agricultural monitoring, enabling automated and scalable analysis of visual data for crop health assessment, yield prediction, and pest surveillance [1]. Among the most persistent and economically damaging threats to rice cultivation are avian pests — particularly Eurasian tree sparrows which target crops during the grain-ripening phase, feeding on rice grains that are nearly ready for harvest

and inflicting significant losses on farmers' incomes and national food security [2–4].

As cultivated areas expand, the scale of these infestations intensifies, and conventional deterrent measures such as scarecrows, reflective tape, and netting have proven increasingly inadequate against large sparrow populations [5], [6]. The deployment of surveillance technologies including CCTV cameras and smartphones offers a compelling foundation for continuous, real-time-capable field monitoring,

enabling automated detection systems to identify pest presence and trigger timely preventive responses [7-9].

When integrated with advanced image analysis algorithms, such systems can substantially reduce dependence on manual labour while supporting more precise and responsive pest management decisions [10], [11].

Despite significant progress in automated visual detection, reliably identifying small, fast-moving pest objects such as sparrows against complex, dynamic agricultural backgrounds remain an open challenge. A broad range of approaches has been explored, from CNN-based classifiers and YOLO-based real-time-capable detectors [7] to training-free saliency-driven algorithms that highlight visually prominent regions without requiring large, annotated datasets [3], [4].

Among these, the Spectral Residual (SR) algorithm has attracted attention for its computational efficiency and applicability to real-time-capable scenarios [10], [12]. Prior studies have demonstrated SR's potential in underwater object detection through combination with fuzzy c-means segmentation, achieving 85% accuracy [13]; in hybrid spatio-temporal video saliency for human action recognition [12]; in roadway salience estimation [14]; and in infrared bird target detection using temporal variation filtering [15].

However, fundamental limitations persist across these implementations. Conventional SR pipelines operate on a single grayscale channel, discarding colour contrast information that is critical for distinguishing small objects from spectrally varied vegetation backgrounds [16-19]. Standard SR saliency maps also highly susceptible to static background textures and lighting transitions, generating excessive false-positive responses in scenes with complex foliage or uneven illumination [20], [21].

The absence of temporal motion integration means that stationary background regions with salient spectral characteristics remain indistinguishable from genuinely moving pest objects, substantially reducing detection reliability in video-based applications [10], [22]. Although lightweight and adaptive saliency models represent an important research direction [23], no existing work has simultaneously addressed colour channel richness, motion awareness, and adaptive thresholding within a unified SR framework for agricultural pest detection [24-26].

In this paper, we present an improved Spectral Residual algorithm for robust, real-time-capable sparrow pest detection in CCTV video footage captured at rice field sites. The proposed method departs from conventional SR. FFT-based spectral analysis is applied independently to each Red, Green, and Blue (RGB) channel preserving the full spectral richness of the agricultural scene and enabling more sensitive saliency responses for objects with distinctive colour signatures [26], [27].

the spatial saliency map produced by the multi-channel SR stage is fused with temporal motion cues extracted via inter-frame difference analysis, suppressing high responses in static textured background regions [15]. adaptive binarization is applied to the fused saliency output to accommodate dynamic lighting variation, followed by Support Vector Machine (SVM) classification with optimised hyperparameters to consolidate the final pest detection decision [14], [28].

The main contributions of this paper are a multi-channel RGB spectral residual framework that independently processes each colour channel via FFT, replacing the conventional single-channel grayscale approach to preserve richer spectral information and improve saliency sensitivity for small pest objects in spectrally complex agricultural environments [26].

Building on this foundation, we introduce a motion-aware saliency integration pipeline that fuses spatial spectral residual maps with temporal motion cues derived from inter-frame difference analysis, effectively suppressing the static-texture false positives [10], [12].

To further consolidate detection reliability, we develop an adaptive binarization and SVM-based classification stage as an end-to-end refinement mechanism, significantly improving robustness under the variable illumination and background dynamics characteristic of open-field CCTV deployments [28-30].

We conduct a rigorous empirical evaluation on a real-world dataset of 136 frames extracted from 2MP Hikvision CCTV footage installed at active rice field sites, benchmarked against the classical SR baseline. The proposed method achieves an accuracy of 98.15%, precision of 91.67%, recall of 98.98%, and F1-score of 95.18%, compared to 74.00%, 70.00%, 65.00%, and 69.00% for the baseline [31], [32].

2. Methods

This study develops and evaluates an improved Spectral Residual algorithm for automated sparrow pest detection in real-world agricultural video footage. The process begins with dataset collection, wherein video footage is captured from a 2MP Hikvision CCTV camera installed at active rice field sites.

This is followed by a preprocessing stage, in which raw video frames undergo extraction, noise filtering, manual segmentation, and pseudo ground truth generation to establish a reliable evaluation reference. The baseline Spectral Residual algorithm is then implemented and analyzed to characterize its inherent limitations under agricultural conditions.

Building on this analysis, the proposed improved algorithm is developed by integrating independent RGB channel processing, temporal motion cue fusion through inter-frame difference analysis, and adaptive binarization to generate more discriminative saliency maps. Support Vector Machine (SVM) classifier with

optimized hyperparameters is employed to consolidate detection decisions, and the full pipeline is evaluated against the baseline using accuracy, precision, recall, and F1-score metrics computed over 136 annotated frames.

2.1 Research Framework

The workflow diagram for this research illustrates the process from dataset collection to final evaluation, including a comparison of the baseline algorithm and the developed algorithm. The research framework provides a systematic and structured workflow that governs all stages of the study, from initial data acquisition through to final performance evaluation, ensuring that the development process remains aligned with the stated research objectives.

As illustrated in Figure 1, the workflow is organised into six sequential stages that collectively form an end-to-end pipeline for agricultural pest detection based on CCTV video imagery [29].

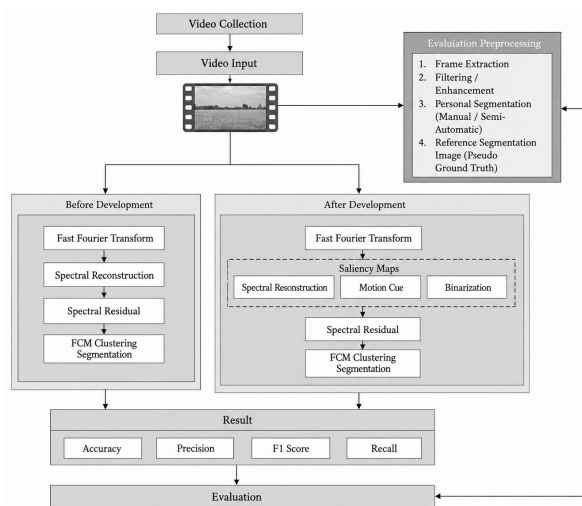


Figure 1. Research Framework

The process begins with video collection, wherein raw footage is acquired from a Hikvision 2MP CCTV camera installed at rice field sites. The collected video serves as the primary input to the system and is selected to represent realistic agricultural conditions. In the second stage, the input video undergoes a preprocessing pipeline comprising frame extraction, noise filtering and image enhancement, manual object segmentation, and the generation of pseudo ground truth masks. This stage establishes the reference data against which both the baseline and the proposed algorithms are objectively evaluated.

The third and fourth stages correspond to the algorithmic development component of the study. The classical Spectral Residual algorithm implemented as the baseline to characterise its performance and expose its limitations under agricultural conditions. The improved algorithm is then developed by incorporating RGB channel processing, temporal motion cue integration via inter-frame difference analysis, and

adaptive binarization, thereby addressing the deficiencies identified in the baseline [30], [31].

In the fifth stage, a Support Vector Machine (SVM) classifier with optimised hyperparameters is applied to the processed saliency outputs to produce final binary detection decisions. The sixth and final stage consists of quantitative accuracy evaluation using precision, recall, F1-score, and overall accuracy metrics computed over 136 annotated frames, followed by a direct comparative analysis between the baseline and the proposed algorithm to validate the contribution of each developmental enhancement [31], [32].

2.2 Dataset

The dataset used in this study was collected from real agricultural environments using a Hikvision EXIR Fixed Mini Bullet Network Camera (2MP), infrared night vision, and motion detection capability [11]. The camera was mounted at a height of approximately two metres above ground level at each monitored site and capture bird movement patterns at a scale consistent with the target pest object size.

Data collection was conducted across 20 rice field sites located within the Padang municipality of West Sumatra, Indonesia, spanning multiple sub-districts including Lubuk Begalung, Kuranji, and Pauh, yielding a total of 360 video recordings. The geographic and environmental diversity of these locations ensures that the dataset reflects a broad range of real-world agricultural scenarios rather than a controlled or laboratory setting.

To support generalisation across different acquisition conditions, supplementary footage was also obtained using smartphone cameras alongside the primary CCTV recordings [9]. This dual-source collection strategy is consistent with approaches reported in the surveillance-based detection literature, where multi-device datasets are used to improve the adaptability of detection models to variations in image resolution, lens distortion, and colour response [22].

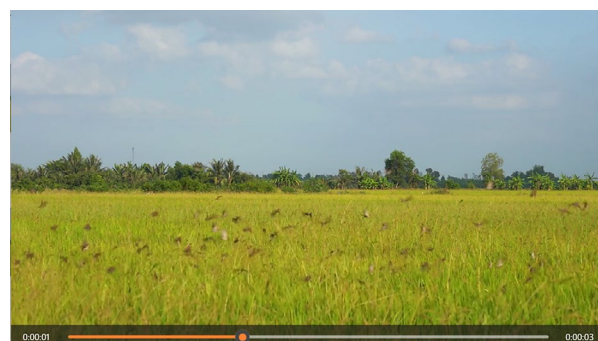


Figure 2. CCTV Pest Detection Results

The dataset was organised into two classes, bird pest present and bird pest absent. From the full collection, 136 frames were extracted and annotated and analysed using MATLAB R2024a on an ASUS laptop equipped with an 11th Gen Intel Core i7-11800H processor and 16 GB of RAM. The dataset is presented as Figure 2,

showing representative CCTV frames captured under the described field conditions.

2.3 Preprocessing for Evaluation

Raw CCTV video footage cannot be used directly for algorithmic evaluation and must first be transformed into a structured, annotated dataset. Figure 3 shows the four-stage preprocessing pipeline applied to all raw CCTV footage prior to algorithmic evaluation. The process begins with frame extraction, in which each video file is read frame-by-frame in MATLAB and exported as individual PNG images at native resolution, preserving full visual detail without additional compression loss [22]. Following this stage, a total of 136 frames are retained as the final evaluation dataset, selected to represent the full range of environmental and lighting conditions observed across the 20 collection sites [1].

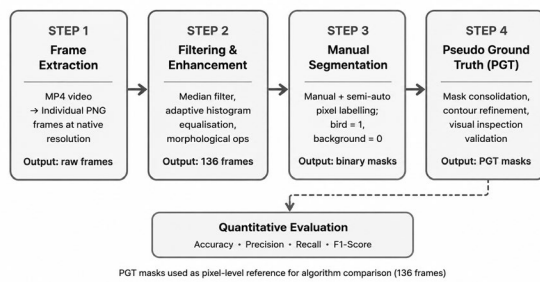


Figure 3. Preprocessing for Evaluation Pipeline

Following frame extraction, a total of 136 frames were retained for evaluation through purposive stratified sampling across the 20 collection sites. Specifically, frames were selected to ensure representation across five key sources of environmental variation observed in the full footage: lighting condition (direct sunlight, overcast, and low-light), background complexity (sparse paddy, dense foliage, and mixed ground textures), time of day (morning, midday, and late afternoon acquisition windows), camera angle and distance variation across the 20 site installations, and the presence or absence of the target bird object.

The study uses a two-tier evaluation design. The broader 600-frame set, created by extracting multiple frames from 360 video files, is visually labeled with a binary tag (bird pest present/absent) and used for SVM training (420 frames) and frame-level testing (180 frames). The 136-frame subset from the same recordings includes additional pixel-level pseudo ground truth masks obtained through the manual segmentation process.

This more labour-intensive pixel-level annotation limited the feasible annotation volume and distinguishes the two datasets. The remaining two stages establish the pixel-level reference required for objective performance measurement. In the manual segmentation stage, bird object masks are produced for each of the 136 frames through a combination of direct manual annotation and semi-automatic thresholding

with researcher correction, assigning a binary label per pixel — one for bird, zero for background [3].

These corrected masks are then consolidated in the final stage to construct the Pseudo Ground Truth (PGT), whereby contours are refined where necessary and each mask is validated through visual inspection. The PGT serves as the gold standard reference against which both the baseline and proposed algorithms are evaluated, with accuracy, precision, recall, and F1-score computed by comparing each algorithm's segmentation output against the corresponding PGT mask at the pixel level [24], [28].

2.4 Baseline Spectral Residual Algorithm

The Spectral Residual (SR) algorithm is a training-free, frequency-domain saliency detection method that identifies perceptually prominent image regions by exploiting statistical irregularities in the log-magnitude spectrum [10], [26]. Its computational efficiency and independence from annotated training data make it well-suited to real-time video analysis, and it is implemented here as the formal baseline against which the proposed improved pipeline is evaluated.

The baseline algorithm operates on a single grayscale representation of each video frame. Each frame is first transformed from the spatial domain into the frequency domain via the two-dimensional Fast Fourier Transform (FFT), yielding a complex spectrum from which the magnitude and phase components are separated. The log-amplitude spectrum is computed as Equation 1.

$$L(f) = \log|\mathcal{F}\{I\}| \quad (1)$$

$\mathcal{F}\{I\}$ denotes the FFT of the input image I . A Gaussian filter h_n is then convolved with $L(f)$ to estimate the average spectral profile of the image, representing its global statistical tendency. The Spectral Residual is obtained by subtracting this smoothed average from the original log-amplitude spectrum as Equation 2.

$$R(f) = L(f) - h_n * L(f) \quad (2)$$

The residual $R(f)$ retains only the frequency components that deviate from the global pattern, which are presumed to be perceptually salient. These are recombined with the original phase spectrum $P(f)$ and mapped back to the spatial domain via the Inverse FFT to produce the final saliency map as Equation 3.

$$S(x) = |\mathcal{F}^{-1}\{\exp(R(f) + P(f))\}|^2 \quad (3)$$

High-intensity regions in $S(x)$ correspond to visually prominent areas in the frame, displayed as white regions against a black background, as shown in Figure 4.

Despite its efficiency, the baseline SR exhibits limitations that are particularly pronounced in open-field agricultural environments. Operating on a single grayscale channel discards chromatic contrast information encoded in the RGB channels, reducing sensitivity to bird objects whose distinction from the background is partly colour-dependent [16],[20].



Figure 4. Spectral Residual Results Before Development

Static regions characterized by intricate spectral textures, such as dense rice foliage, frequently generate high-clarity responses that are indistinguishable from genuine detection [12], [20]. Concurrently, the absence of temporal processing renders inter-frame motion information entirely unutilized. These limitations are reflected in the baseline evaluation results: accuracy 74.00%, precision 70.00%, recall 65.00%, and F1-score 69.00%, establishing the performance floor that the proposed method is designed to surpass [33].

2.5 Proposed Improved Spectral Residual Algorithm

The proposed method extends the baseline SR pipeline through three sequential developments, multi-channel RGB spectral processing, temporal motion cue integration, and adaptive binarization with SVM classification, as illustrated in Figure 5. Each stage directly addresses a specific limitation of the baseline while preserving the training-free, computationally lightweight character of the original algorithm.

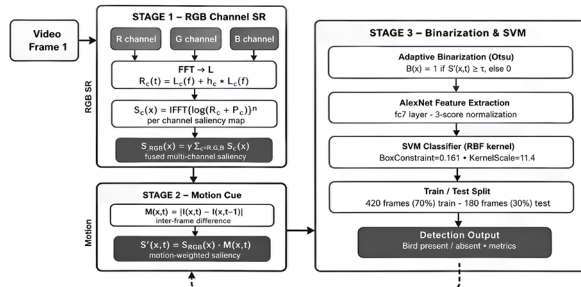


Figure 5. Proposed Improved Spectral Residual Algorithm Pipeline

In Stage 1, the single grayscale input of the baseline is replaced by independent FFT-based SR processing of each colour channel $c \in \{R, G, B\}$. The per-channel spectral residual and saliency map are computed as $R_c(f) = L_c(f) - h_n * L_c(f)$ and $S_c(x) = |\mathcal{F}^{-1}\{\exp(R_c(f) + P_c(f))\}|^2$, after which the three maps are fused by channel averaging into a unified multi-channel saliency map $S_{RGB}(x) = (1/3) \sum S_c(x)$.

In Stage 2, temporal motion cues are integrated to suppress false positives arising from static but texturally complex backgrounds. An inter-frame difference map $M(x, t) = |I(x, t) - I(x, t-1)|$ is computed and fused with $S_{RGB}(x)$ through element-wise multiplication, yielding $S'(x, t) = S_{RGB}(x) \odot M(x, t)$. This reinforces responses corresponding to genuine object movement

while attenuating static background regions that produce spurious saliency activations [12], [15].

In Stage 3, the fused saliency map is binarised using an adaptive Otsu threshold τ , producing the detection mask $B(x) = 1$ if $S'(x, t) \geq \tau$, else 0, which adapts to per-frame illumination variation rather than applying a fixed global threshold [33],[34]. Features are then extracted from the fc7 layer of AlexNet via transfer learning, normalised using z-score standardisation, and classified by an RBF-kernel SVM [35] with optimised hyperparameters (BoxConstraint = 0.161, KernelScale = 11.397, objective function = 0.0046), trained on 420 frames and evaluated on 180 frames from the 600-frame total dataset [28].

2.6 Accuracy Evaluation

Performance evaluation is conducted at the pixel level by comparing each algorithm's binary detection output against the corresponding Pseudo Ground Truth mask across all 136 evaluation frames. Each pixel is classified into one of four categories, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), where TP denotes a correctly detected bird pixel, TN a correctly identified background pixel, FP a background pixel incorrectly labelled as bird, and FN a bird pixel missed by the algorithm.

Accuracy alone is insufficient as the sole evaluation criterion in this context because rice field imagery is characterised by a significant class imbalance which causes accuracy to be disproportionately inflated by correct background classifications [30].

Precision captures the algorithm's ability to avoid false alarms, recall measures its ability to detect all actual bird instances without omission, and the F1-score provides a single balanced measure of both. Together, these four metrics provide a comprehensive and bias-resistant assessment of detection performance [28].

3. Results and Discussions

This section presents the empirical findings obtained from implementing and evaluating both the baseline and proposed Spectral Residual algorithms on the 136-frame CCTV dataset collected from active rice field sites.

Results are reported across four stages: the preprocessing output that constitutes the evaluation dataset, a qualitative visual comparison of saliency map outputs before and after algorithmic development, a quantitative performance evaluation using accuracy, precision, recall, and F1-score metrics derived from confusion matrix analysis, and a discussion that interprets these findings in relation to the research gap identified in the Introduction and the existing literature.

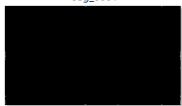
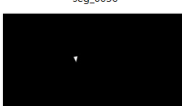
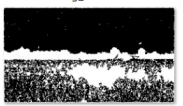
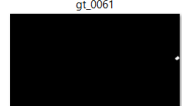
All experiments were conducted in MATLAB R2024a on an ASUS laptop equipped with an 11th Gen Intel Core i7-11800H processor and 16 GB of RAM, with the proposed pipeline evaluated against the classical SR

baseline under identical data and testing conditions to ensure a fair and reproducible comparison.

3.1 Preprocessing Output

The preprocessing pipeline described in Section 2.3 was applied to all video footage, yielding a structured evaluation dataset of 136 annotated frames. Representative outputs from each preprocessing stage are presented in Table 1, which displays the original frame, filtered image, manual segmentation mask, and corresponding pseudo ground truth for a selection of frames across the dataset.

Table 1. Representative Preprocessing Output Samples

Frame	Original Frame	Filtered Output	Segmentation Mask	Pseudo Ground Truth (PGT)
Frame 000001				
	frame_000001	bg_000001	seg_0001	gt_0001
Frame 000002				
	frame_000002	bg_000002	seg_0002	gt_0002
Frame 000036				
	frame_000036	bg_000036	seg_0036	gt_0036
Frame 000100				
	frame_000100	bg_000061	gt_0061	gt_0061
Frame 000136				
	frame_000136	seg_0100	gt_0100	gt_0100

The segmentation results exhibit natural variability across the dataset. In frames captured under low-contrast or complex background conditions some degree of over-segmentation is observable, where non-object regions are incorrectly included in the foreground mask. Conversely, a small number of frames exhibit under-segmentation, where parts of the bird object are not fully captured due to poor contrast between the target and background [28].

These observations are consistent with the inherent challenges of annotating small, fast-moving objects in open-field CCTV footage. The completed pseudo ground truth masks, derived from the corrected segmentation outputs and validated through visual inspection, provide a reliable pixel-level reference for the quantitative evaluations.

3.2 Qualitative Results: Saliency Map Comparison

Figures 6 and 7 present representative saliency map outputs from the baseline and proposed algorithms respectively.

As shown in Figure 6, the baseline Spectral Residual algorithm produces saliency maps that identifying

Following frame extraction, median filtering and adaptive histogram equalisation produced noticeably cleaner binary representations, with object candidate regions more sharply defined and spurious background noise substantially reduced. Frames 36 and 61 demonstrate that filtering effectively preserves relevant object areas while suppressing isolated noise pixels. The subsequent manual and semi-automatic segmentation stage successfully produced binary masks for all 136 frames, with bird pixels assigned a value of one and background pixels assigned zero.

regions of general visual prominence, exhibit several characteristic deficiencies in the agricultural context. The detected regions appear as diffuse, irregularly shaped white blobs with poorly defined boundaries.



Figure 6. Spectral Residual Results Before Development

More critically, static background regions, particularly dense rice foliage and ground surface textures, generate consistently high saliency responses that are visually indistinguishable from genuine bird detections.

Figure 7 demonstrates a marked qualitative improvement in the saliency maps produced by the proposed algorithm. The detected regions are visibly

sharper, more compact, and better localised to areas corresponding to actual bird movement, with the surrounding background suppressed to near-black.

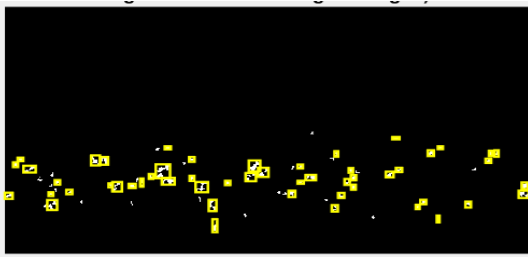


Figure 7. Spectral Residual Results After Development

This improvement is directly attributable to the three-stage pipeline: the RGB channel processing recovers chromatic contrast suppressed by grayscale conversion, the inter-frame motion cue integration eliminates static background activations that do not correspond to genuine movement, and the adaptive binarization produces cleaner binary boundaries that better conform to the shape and position of the target objects [12],[33]. The distinction between bird pixels and background pixels is substantially clearer in the post-development outputs, with fewer scattered false detections visible across the frame.

Table 2. Qualitative Comparison of Saliency Map Outputs: Baseline SR vs. Proposed SR

Frame	Baseline SR Output (Before Development)	Proposed SR Output (After Development)
Frame 36		

The frame-level comparison in Table 2 reinforces these observations consistently across the evaluation dataset. In frames where the baseline exhibits over-segmentation, the proposed algorithm confines its detections to regions of genuine movement, demonstrating a clear reduction in false-positive area.

3.3 Quantitative Performance Evaluation

Quantitative evaluation was conducted at two levels. At the pixel level, 136 representative frames were assessed against pseudo ground-truth (PGT) binary masks using four standard metrics: accuracy, precision, recall, and F1-score. At the frame classification level, a Support Vector Machine (SVM) with a radial basis function (RBF) kernel was trained on AlexNet fc7 features [36] extracted from 600 labelled frames (420 training, 180 testing). To ensure optimal classification performance, SVM hyperparameters were tuned through Bayesian optimisation [37] over 30 iterations, with the results summarised in Table 3.

The optimisation converged to a BoxConstraint of 0.161 and a KernelScale of 11.397, with a minimum objective function value of 0.0046 indicating that the SVM classifier reached a highly stable and well-generalised solution within the defined search space.

The corresponding confusion matrix on the 180-frame test set is shown in Figure 8.

Table 3. SVM Hyperparameter Optimisation Results

Parameter	Search Range	Optimal Value
BoxConstraint	[0.001, 1000]	0.161
KernelScale	[0.001, 1000]	11.397
Minimum Objective Value	—	0.0046
Iterations Completed	—	30

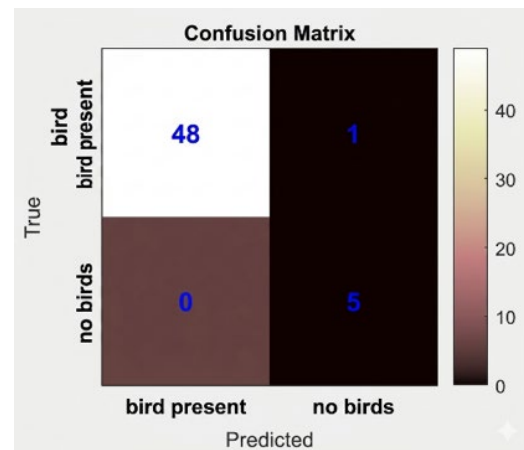


Figure 8. Confusion Matrix of the After Development

The confusion matrix confirms the strong discriminative capability of the proposed pipeline. The aggregated pixel-level and frame-level performance metrics across both the baseline and proposed methods are summarised in Table 3. Given that rice field imagery contains a substantially larger proportion of background pixels relative to bird-object pixels, accuracy alone is an insufficient indicator of detection quality. Precision, recall, and F1-score are therefore reported to provide a balanced assessment that accounts for class imbalance.

Table 4. Performance Metrics

Metric	Before Development	After Development
Accuracy	74.00%	98.15%
Precision	70.00%	91.67%
Recall	65.00%	98.98%
F1-score	69.00%	95.18%

As reported in Table 4, the proposed method achieves an accuracy of 98.15%, precision of 91.67%, recall of 98.98%, and F1-score of 95.18%, representing substantial gains over the baseline SR, which scored 74.00%, 70.00%, 65.00%, and 69.00% on the respective metrics. The notably high recall of 98.98% confirms that the temporal motion integration stage successfully suppresses static background saliency, ensuring that nearly all true bird-object pixels are captured.

3.4 Discussion

The most diagnostically informative result is the recall improvement from 65.00% to 98.98%. In the baseline SR, saliency is derived from a single grayscale channel without temporal context, causing static high-contrast background regions — such as rippling water surfaces and sunlit paddy textures — to generate persistent salient responses that compete with and suppress

genuine bird detections. The temporal motion term $M(x,t) = |I(x,t) - I(x,t-1)|$ selectively amplifies only those salient regions that coincide with inter-frame change, thereby eliminating background-induced false negatives at the source.

The precision gain from 70.00% to 91.67% is attributable to adaptive Otsu binarisation, which derives a frame-specific threshold from the bimodal intensity distribution of the fused saliency map. The resulting F1-score of 95.18% demonstrates that both objectives are achieved simultaneously, confirming that the three pipeline stages address complementary failure modes of the baseline method.

The performance gap between the baseline and proposed SR reflects the architectural limitation of single-channel, threshold-fixed saliency detection in complex agricultural scenes. Saliency-based approaches for grain pest detection, such as the cascaded aggregation network reported by Yu et al. [3], have achieved weighted F-measures of 76.4–91.0%, while YOLO-based frameworks applied to field pest surveillance report F1-scores in the range of 79–83% [7]. CNN-based bird detection systems, including MobileNetV2 architectures, attain accuracy levels around 80–92% but do not incorporate temporal motion discrimination [17].

Methods employing temporal variation filtering for infrared bird target detection have demonstrated the value of temporal cues in suppressing background clutter, though their application has been limited to infrared imagery rather than standard CCTV footage [15]. It is important to position the proposed method within the broader context of modern object detection, where lightweight deep learning architectures have become the dominant paradigm. Models such as YOLOv8-nano (~3.2M parameters), MobileNetV2 [17], and EfficientDet-Lite0 (~3.4M parameters) represent the current frontier of edge-deployable detectors and, when trained on sufficiently large and diverse annotated datasets, consistently achieve detection performance that generalized saliency-based methods cannot match in absolute terms.

YOLO-based frameworks applied to field pest surveillance, for instance, have reported F1-scores in the range of 79–83% on large multi-class datasets[6], and with domain-specific fine-tuning, such models can approach or exceed the accuracy figures reported in the present study.

However, this comparison is not straightforwardly in favour of deep learning when deployment context is considered alongside raw performance. Training a lightweight deep learning detector requires a substantial labeled dataset — typically on the order of thousands of annotated bounding boxes or segmentation masks — a dedicated GPU for training, and ongoing maintenance as environmental conditions shift.

In agricultural pest surveillance contexts such as open paddy fields in developing regions, precisely these

resources are in chronic short supply: annotated bird detection datasets are scarce, GPU infrastructure is typically unavailable at field scale, and model retraining following seasonal or site-specific variation is logistically demanding. The proposed method, by contrast, requires no training data, no learnable weights, no GPU infrastructure, and carries a near-zero memory footprint beyond the FFT buffer — properties that translate directly into deployment feasibility in resource-constrained agricultural environments.

The computational profile of the two paradigms also differs in kind rather than merely in degree. Lightweight deep learning models incur a fixed but non-trivial one-time training cost (GPU hours, annotation labour, storage of weight matrices) and a per-inference cost dominated by large matrix multiplications. The proposed pipeline's per-frame cost is dominated by three FFT passes ($O(N \log N)$ each), one inter-frame subtraction ($O(N)$), and one Otsu thresholding operation ($O(N)$) — operations for which highly optimized CPU implementations exist in standard signal processing toolchains.

The appropriate comparison is therefore not which approach achieves higher accuracy on a data-rich benchmark, but which offers a viable and maintainable detection capability under the real-world constraints of smallholder rice farming infrastructure. The proposed method is positioned as a practical, deployable baseline for precisely such contexts, and is not intended to supersede deep learning detectors where sufficient labeled data and compute are available.

A transparent limitation of the present study is the evaluation set size of 136 annotated frames. While this scale is consistent with pixel-level saliency evaluation practice in the agricultural detection literature and was drawn from a geographically diverse pool of 360 recordings across 20 active field sites, it remains small relative to the training corpora typically associated with deep learning baselines.

The dataset does not capture seasonal variation across the rice growth cycle, nor conditions beyond the Padang municipality of West Sumatra, and the generalizability of the reported metrics to other agro-climatic zones, camera configurations, or sparrow population densities cannot be assumed without further validation. Cross-regional replication with a substantially larger annotated set is accordingly identified as the primary dataset-related priority for future work.

The current implementation operates in offline batch mode, and real-time deployment on edge hardware would require architectural optimisation to meet processing latency constraints at field scale.

A brief theoretical analysis of edge deployment feasibility is warranted, given that real-world agricultural monitoring systems are typically constrained to standard embedded hardware rather than laboratory workstations. The two computationally distinct components of the proposed pipeline — the

FFT-based saliency stage and the AlexNet feature extraction stage — exhibit markedly different edge deployment profiles and must be assessed separately.

The FFT-based saliency stage, comprising three per-channel spectral residual computations, inter-frame differencing, channel averaging, and adaptive Otsu thresholding, is theoretically well-suited to standard edge devices. FFT operations at agricultural CCTV resolution (typically 1920×1080, commonly downsampled to 640×480 for processing) are tractable on ARM Cortex-A processors using SIMD-accelerated libraries such as FFTW or ARM Ne10, which exploit NEON vector instructions available on devices including the Raspberry Pi 4 and NVIDIA Jetson Nano.

Memory requirements for the FFT buffer at 640×480 resolution remain well below 10MB — comfortably within the RAM envelope of all commonly deployed edge platforms. Inter-frame differencing and Otsu thresholding are $O(N)$ operations and contribute negligible additional cost. The saliency stage of the pipeline may therefore be considered a strong candidate for edge deployment without architectural modification.

The AlexNet feature extraction stage presents a more significant constraint. AlexNet carries approximately 60 million parameters (~240MB in single-precision floating point), and inference through the full network to the fc7 layer is computationally demanding on CPU-only edge hardware — estimated at approximately 1–2 seconds per frame on a Raspberry Pi 4 in an unoptimised configuration, which precludes interactive operation.

On GPU-enabled edge platforms such as the NVIDIA Jetson Nano (472 GFLOPS, 4GB LPDDR4), AlexNet inference is substantially more tractable and compatible with near-real-time throughput. The most pragmatic resolution to this bottleneck, however, is architectural substitution: replacing AlexNet as the feature backbone with a lightweight convolutional model such as MobileNetV2 (~3.4M parameters, ~14MB) or SqueezeNet (~1.2M parameters, ~5MB) would reduce the feature extraction footprint by one to two orders of magnitude while preserving the overall SVM classification architecture. This substitution is identified as the primary engineering priority for edge-targeted implementations of the proposed pipeline, and its effect on classification accuracy relative to AlexNet fc7 features is a concrete direction for future work.

The generalizability of the proposed method is bounded by the conditions under which the dataset was collected. All 136 evaluation frames originate from a single tropical equatorial environment in Padang, West Sumatra, where background textures, illumination profiles, and spectral properties are specific to Indonesian paddy fields during the grain-ripening phase.

The algorithm was furthermore developed and validated exclusively for Eurasian tree sparrows (*Passer montanus*); other economically significant avian rice

pests — including Red-billed Quelea in sub-Saharan Africa and various weaver species — differ substantially in body size, plumage, and flight behavior, and the trained SVM classifier cannot be assumed to transfer to these targets without revalidation.

Additional constraints arise from camera and seasonal dependency. The inter-frame difference stage is sensitive to compression artifacts that vary across camera models and encoding settings, while night-mode infrared operation collapses RGB channel diversity into near-monochromatic output, eliminating the primary advantage of multi-channel processing over the grayscale baseline.

Background structure also changes dramatically across the rice growth cycle, from sparse seedling beds to dense vegetative canopy, introducing textures absent from the present dataset. These boundaries collectively define the conditions under which the reported metrics are valid and constitute a concrete agenda for future cross-regional, multi-species, and full growth-cycle validation

Future work should address cross-regional validation, extension to multi-species pest scenarios, and investigation of lightweight convolutional feature extractors as alternatives to AlexNet.

4. Conclusions

This study proposed an improved Spectral Residual algorithm for automated sparrow pest detection in paddy field CCTV footage, extending the baseline method through three targeted enhancements: independent RGB channel saliency fusion, inter-frame motion cue integration, and adaptive Otsu binarisation followed by AlexNet-SVM classification. Evaluated on 136 annotated frames against pseudo ground-truth masks, the proposed pipeline achieved an accuracy of 98.15%, precision of 91.67%, recall of 98.98%, and F1-score of 95.18%, compared to 74.00%, 70.00%, 65.00%, and 69.00% for the baseline SR. These results confirm the viability of the proposed approach as a practical, training-free saliency backbone for early pest detection systems, with future work directed toward cross-regional validation, multi-species extension, and real-time edge deployment.

References

- [1] E. F. Duque-Vazquez, R. E. Sanchez-Yanez, N. Saldaña-Robles, M. F. León-Galván, and J. Cepeda-Negrete, “HeLa cell segmentation using digital image processing,” *Heliyon*, vol. 10, no. 5, p. e26520, Mar. 2024, doi: [10.1016/j.heliyon.2024.e26520](https://doi.org/10.1016/j.heliyon.2024.e26520).
- [2] Y. Wang et al., “WBNet: Weakly-supervised salient object detection via scribble and pseudo-background priors,” *Pattern Recogn.*, vol. 154, no. C, 2024, doi: [10.1016/j.patcog.2024.110579](https://doi.org/10.1016/j.patcog.2024.110579).
- [3] J. Yu, S. Chen, N. Liu, F. Zhai, and Q. Pan, “Cascaded Aggregation Convolution Network for Salient Grain Pests Detection,” *Insects*, vol. 15, no. 7, 2024, doi: [10.3390/insects15070557](https://doi.org/10.3390/insects15070557).
- [4] I. Ullah et al., “CMGNet: Context-aware middle-layer guidance network for salient object detection,” *Journal of King Saud*

- University - Computer and Information Sciences, vol. 36, no. 1, p. 101838, 2024, doi: [10.1016/j.jksuci.2023.101838](https://doi.org/10.1016/j.jksuci.2023.101838).
- [5] C. Zhang, K. Dong, K. Aihara, L. Chen, and S. Zhang, "STAMarker: determining spatial domain-specific variable genes with saliency maps in deep learning," *Nucleic Acids Res.*, vol. 51, no. 20, pp. e103–e103, Nov. 2023, doi: [10.1093/nar/gkad801](https://doi.org/10.1093/nar/gkad801).
 - [6] Y. Zhang, Z. Zhang, P. Zhang, and M. Xu, "Salient object detection for RGBD video via spatial interaction and depth-based boundary refinement," *Complex & Intelligent Systems*, vol. 9, no. 6, pp. 6343–6358, 2023, doi: [10.1007/s40747-023-01072-w](https://doi.org/10.1007/s40747-023-01072-w).
 - [7] E. C. Tetila et al., "YOLO performance analysis for real-time detection of soybean pests," *Smart Agricultural Technology*, vol. 7, p. 100405, 2024, doi: [10.1016/j.atech.2024.100405](https://doi.org/10.1016/j.atech.2024.100405).
 - [8] J. A. Figueroa-Alvarez et al., "Insectivorous birds and potential pest control services: An occupancy study of functional groups in a coffee landscape in Oaxaca, Mexico," *Perspect. Ecol. Conserv.*, vol. 22, no. 4, pp. 331–341, 2024, doi: [10.1016/j.pecon.2024.09.002](https://doi.org/10.1016/j.pecon.2024.09.002).
 - [9] A. Raza, M. R. Al Nasar, E. S. Hanandeh, R. A. Zitar, A. Y. Nasereddin, and L. Abualigah, "A Novel Methodology for Human Kinematics Motion Detection Based on Smartphones Sensor Data Using Artificial Intelligence," 2023, doi: [10.3390/technologies11020055](https://doi.org/10.3390/technologies11020055).
 - [10] V. C. Warad and R. Fatima, "Video saliency-recognition by applying custom spatio temporal fusion technique," *IAES International Journal of Artificial Intelligence*, vol. 13, no. 1, pp. 82–91, 2024, doi: [10.11591/ijai.v13.i1.pp82-91](https://doi.org/10.11591/ijai.v13.i1.pp82-91).
 - [11] J. Hentati-Sundberg et al., "Seabird surveillance: combining CCTV and artificial intelligence for monitoring and research," *Remote Sens. Ecol. Conserv.*, vol. 9, no. 4, pp. 568–581, Aug. 2023, doi: [10.1002/rse2.329](https://doi.org/10.1002/rse2.329).
 - [12] A. A. Gharahbagh, V. Hajhashemi, M. C. Ferreira, J. J. M. Machado, and J. M. R. S. Tavares, "Hybrid time-spatial video saliency detection method to enhance human action recognition systems," *Multimed. Tools Appl.*, vol. 83, no. 30, pp. 74053–74073, 2024, doi: [10.1007/s11042-024-18126-x](https://doi.org/10.1007/s11042-024-18126-x).
 - [13] Y. Wen, Y. Wang, H. Wang, W. Shi, and W. Cao, "Co-salient object detection with iterative purification and predictive optimization," *Virtual Reality & Intelligent Hardware*, vol. 6, no. 5, pp. 396–407, 2024, doi: [10.1016/j.vrih.2024.06.002](https://doi.org/10.1016/j.vrih.2024.06.002).
 - [14] Z. Berger Haladova, M. Zrubec, and Z. Cernekova, "A method for estimating roadway billboard salience," in *ACM Symposium on Applied Perception 2024*, in SAP '24. New York, NY, USA: Association for Computing Machinery, 2024, doi: [10.1145/3675231.3675238](https://doi.org/10.1145/3675231.3675238).
 - [15] F. Zhao, R. Wei, Y. Chao, S. Shao, and C. Jing, "Infrared Bird Target Detection Based on Temporal Variation Filtering and a Gaussian Heat-Map Perception Network," 2022, doi: [10.3390/app12115679](https://doi.org/10.3390/app12115679).
 - [16] C. Gwon and S. C. Howell, "ODSmoothGrad: Generating Saliency Maps for Object Detectors," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 3686–3690, doi: [10.1109/CVPRW59228.2023.00376](https://doi.org/10.1109/CVPRW59228.2023.00376).
 - [17] S. Andika Maulana, S. Husna Batubara, Y. Permata Putri Pasaribu, H. Syahputra, and F. Ramadhani, "Deteksi Burung Menggunakan Convolutional Neural Network (Cnn) Dengan Model Arsitektur Mobilenetv2," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 4, pp. 6108–6114, 2024, doi: [10.36040/jati.v8i4.10126](https://doi.org/10.36040/jati.v8i4.10126).
 - [18] M. Olague, G. Olague, R. Pineda, and G. Ibarra-Vazquez, "Salient object detection dataset with adversarial attacks for genetic programming and neural networks," *Data Brief*, vol. 57, p. 111043, 2024, doi: [10.1016/j.dib.2024.111043](https://doi.org/10.1016/j.dib.2024.111043).
 - [19] Y. Zhu and T. H. Falk, "Spectral-temporal saliency masks and modulation tensorgrams for generalizable COVID-19 detection," *Comput. Speech Lang.*, vol. 86, p. 101620, 2024, doi: [10.1016/j.csl.2024.101620](https://doi.org/10.1016/j.csl.2024.101620).
 - [20] A.-C. Woerl, J. Disselhoff, and M. Wand, "Initialization Noise in Image Gradients and Saliency Maps," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 1766–1775, doi: [10.1109/CVPR52729.2023.00176](https://doi.org/10.1109/CVPR52729.2023.00176).
 - [21] D. A. Septian and A. Suhendar, "Klasifikasi Penyakit pada Daun Kentang Menggunakan Pengolahan Citra dengan Metode Convolutional Neural Network (CNN)," *Jurnal teknologi dan sains*, pp. 1017–1025, 2024, doi: [10.51401/jinteks.v6i4.4880](https://doi.org/10.51401/jinteks.v6i4.4880).
 - [22] Z.-W. Sun et al., "FBD-SV-2024: Flying Bird Object Detection Dataset in Surveillance Video," *Sci. Data*, vol. 12, no. 1, p. 530, 2025, doi: [10.1038/s41597-025-04872-6](https://doi.org/10.1038/s41597-025-04872-6).
 - [23] N.-T. Thu, H. N. Tran, Md. D. Hossain, and E.-N. Huh, "LightSOD: Towards lightweight and efficient network for salient object detection," *Computer Vision and Image Understanding*, vol. 249, p. 104148, 2024, doi: [10.1016/j.cviu.2024.104148](https://doi.org/10.1016/j.cviu.2024.104148).
 - [24] Y. A. D. Djilali, K. McGuinness, and N. O'Connor, "Learning Saliency From Fixations," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 382–392, doi: [10.1109/WACV57701.2024.00045](https://doi.org/10.1109/WACV57701.2024.00045).
 - [25] Z. Li, X. Pang, Y. Jiang, and Y. Wang, "RealFuVSR: Feature Enhanced Real-World Video Super-Resolution," *Virtual Reality & Intelligent Hardware*, vol. 5, no. 6, pp. 523–537, 2023, doi: [10.1016/j.vrih.2023.06.006](https://doi.org/10.1016/j.vrih.2023.06.006).
 - [26] P. Liu, T. Xu, G. Shi, J. Xu, H. Chen, and J. Li, "Spectrum-Oriented Point-Supervised Saliency Detector for Hyperspectral Images," *IEEE Trans. Instrum. Meas.*, vol. 74, no. 8, pp. 1–15, 2025, doi: [10.1109/TIM.2025.3527488](https://doi.org/10.1109/TIM.2025.3527488).
 - [27] L. Li et al., "An Effective Infrared and Visible Image Fusion Approach via Rolling Guidance Filtering and Gradient Saliency Map," 2023, doi: [10.3390/rs15102486](https://doi.org/10.3390/rs15102486).
 - [28] K. Szczepankiewicz et al., "Ground truth based comparison of saliency maps algorithms," *Sci. Rep.*, vol. 13, no. 1, p. 16887, 2023, doi: [10.1038/s41598-023-42946-w](https://doi.org/10.1038/s41598-023-42946-w).
 - [29] G. Cartella, V. Cuculo, M. Cornia, M. Papasidero, F. Ruozi, and R. Cucchiara, "Pixels of Faith: Exploiting Visual Saliency to Detect Religious Image Manipulation," in *Computer Vision – ECCV 2024 Workshops: Milan, Italy, September 29–October 4, 2024, Proceedings, Part VI*, Berlin, Heidelberg: Springer-Verlag, 2024, pp. 229–245, doi: [10.1007/978-3-031-91572-7_14](https://doi.org/10.1007/978-3-031-91572-7_14).
 - [30] N. Tomar, S. Chandel, and G. Bhatnagar, "A visual attention-based algorithm for brain tumor detection using an on-center saliency map and a superpixel-based framework," *Healthcare Analytics*, vol. 5, p. 100323, 2024, doi: [10.1016/j.health.2024.100323](https://doi.org/10.1016/j.health.2024.100323).
 - [31] A. Moskalenko et al., "AIM 2024 Challenge on Video Saliency Prediction: Methods and Results," *Lecture Notes in Computer Science*, vol. 15632 LNCS, pp. 178–194, 2025, doi: [10.1007/978-3-031-91856-8_11](https://doi.org/10.1007/978-3-031-91856-8_11).
 - [32] H. Lee, J. Kim, and J. H. Park, "Power System Transient Stability Prediction Using Convolution Neural Network and Saliency map," in *2023 IEEE PES Innovative Smart Grid Technologies - Asia (ISGT Asia)*, 2023, pp. 1–5, doi: [10.1109/ISGTAsia54891.2023.10372670](https://doi.org/10.1109/ISGTAsia54891.2023.10372670).
 - [33] M. N. Darpito, K. Firdausy, and A. Fadil, "Comparative Evaluation of Thresholding Methods for Optimized Digital Document Parsing Accuracy," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 10, no. 2 SE-Computer Vision and Pattern Recognition, Apr. 2026, doi: [10.29207/resti.v10i2.6982](https://doi.org/10.29207/resti.v10i2.6982).
 - [34] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Trans. Syst. Man Cybern.*, vol. 9, no. 1, pp. 62–66, 1979, doi: [10.1109/TSMC.1979.4310076](https://doi.org/10.1109/TSMC.1979.4310076).
 - [35] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, doi: [10.1007/BF00994018](https://doi.org/10.1007/BF00994018).
 - [36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017, doi: [10.1145/3065386](https://doi.org/10.1145/3065386).
 - [37] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," *Adv. Neural Inf. Process. Syst.*, vol. 4, pp. 2951–2959, 2012, doi: [10.48550/arXiv.1206.2944](https://doi.org/10.48550/arXiv.1206.2944).