

# DETEKSI AKTIVITAS BUANG SAMPAH PADA VIDEO CCTV MENGGUNAKAN MOVINET-A0 DENGAN DETEKSI MANUSIA BERBASIS YOLOV11

Raden Ayu Citra Dewi<sup>1</sup>, R. D. Kusumanto<sup>2</sup>, Nyayu Latifah Husni<sup>3\*</sup>

<sup>1,2,3</sup>Department of Electrical Engineering, Politeknik Negeri Sriwijaya; Jl. Srijaya Negara, Bukit Besar, Palembang, South Sumatra, 30139; (0711) 353414/(0711) 355918

## Keywords:

Deteksi Aktivitas Buang Sampah; MoViNet-A0; YOLOv11; Deep Learning; Video CCTV

## Correspondent Email:

nyayu\_latifah@polsri.ac.id

**Abstrak.** Pembuangan sampah sembarangan di badan air perkotaan seperti Sungai Sekanak, Palembang, masih sulit diawasi secara manual melalui rekaman CCTV konvensional. Penelitian ini mengusulkan sistem deteksi aktivitas membuang sampah yang menggabungkan YOLOv11 sebagai detektor manusia untuk mengekstraksi *Region of Interest* (ROI) dan MoViNet-A0 sebagai model pengenalan aksi video untuk mengklasifikasikan aktivitas menjadi Buang Sampah dan Normal. Dataset penelitian terdiri dari 1.413 klip video CCTV (548 Buang Sampah dan 865 Normal) yang direkam pada siang hari di kawasan Sungai Sekanak dan dibagi dengan rasio 70:15:15 untuk *training*, *validation*, dan *testing*. Hasil pengujian pada data uji menunjukkan model mencapai *accuracy* sebesar 92,92%, dengan *precision* 92,41%, *recall* 89,02%, dan *F1-score* 90,68% pada kelas Buang Sampah, serta *precision* 93,23%, *recall* 95,38%, dan *F1-score* 94,30% pada kelas Normal. Pengujian *real-time* langsung menggunakan kamera CCTV pada 10 sampel aktivitas menghasilkan akurasi 70%, lebih rendah dibandingkan akurasi pelatihan, dengan kesalahan klasifikasi yang umumnya terjadi pada objek berjarak jauh dari kamera dan tertutup sebagian oleh vegetasi. Hasil ini menunjukkan bahwa kombinasi YOLOv11 dan MoViNet-A0 mampu mendeteksi aktivitas membuang sampah dengan akurasi tinggi pada data terkurasi, namun memerlukan penyempurnaan lebih lanjut agar lebih andal pada kondisi pengawasan nyata yang lebih bervariasi.



Copyright © 2026 The Author(s), Published by Jurnal Informatika dan Teknik Elektro Terapan. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

**Abstract.** Illegal littering along urban water bodies such as Sungai Sekanak, Palembang, remains difficult to monitor manually through conventional CCTV footage. This study proposes a littering activity detection system that combines YOLOv11 as a human detector to extract the human Region of Interest (ROI) and MoViNet-A0 as a video action recognition model to classify activities into Littering and Normal. The dataset consists of 1,413 CCTV video clips (548 Littering and 865 Normal) recorded during daytime around the Sekanak River area, split with a 70:15:15 ratio for training, validation, and testing. Evaluation on the test set shows that the model achieves an accuracy of 92.92%, with a precision of 92.41%, recall of 89.02%, and F1-score of 90.68% for the Littering class, and a precision of 93.23%, recall of 95.38%, and F1-score of 94.30% for the Normal class. A real-time field test using a live CCTV camera on 10 activity samples achieved 70% accuracy, lower than the training accuracy, with misclassifications mostly occurring when objects were far from the camera or partially occluded by vegetation. These results indicate that the combination of YOLOv11 and MoViNet-A0 can accurately detect littering activity on curated data, but requires further refinement to

*achieve reliable performance under more variable real-world surveillance conditions.*

## 1. PENDAHULUAN

Salah satu masalah lingkungan utama di daerah perkotaan adalah pembuangan sampah sembarangan, yang menyebabkan pencemaran air, kerusakan ekosistem, dan bahaya bagi kesehatan masyarakat. Meskipun ada peraturan dan kampanye kesadaran publik, pembuangan limbah yang tidak benar di sungai dan ruang publik masih terjadi di banyak negara berkembang, termasuk Indonesia. Metode pemantauan konvensional tidak efisien untuk pengawasan dan intervensi cepat karena bergantung pada observasi manual atau inspeksi pasca-kejadian. Akibatnya, ada permintaan yang meningkat untuk sistem pengawasan otomatis yang menggunakan metode *deep learning* untuk secara otomatis mendeteksi pembuangan sampah dari rekaman CCTV [1], [2].

Melalui integrasi pengenalan aksi video dan deteksi objek, kemajuan terbaru dalam visi komputer telah secara signifikan meningkatkan sistem pengawasan otomatis. Dalam video pengawasan, teknik deteksi objek biasanya digunakan untuk melokalisasi subjek manusia. Di sisi lain, model pengenalan aksi video menganalisis informasi spasial dan temporal untuk mengenali aktivitas manusia. Mengekstraksi *Regions of Interest* (ROI) manusia memungkinkan model pengenalan aksi untuk fokus pada informasi visual yang paling relevan, sehingga mengurangi gangguan latar belakang dan meningkatkan efisiensi komputasi. Ini terjadi karena metode ini memungkinkan model untuk fokus pada informasi visual yang paling relevan daripada memproses seluruh bingkai video. Selain itu, perkembangan terbaru dalam arsitektur pengenalan aksi video ringan, seperti MoViNet, telah menunjukkan bahwa akurasi pengenalan yang tinggi dapat dicapai sambil mempertahankan kompleksitas komputasi yang rendah. Ini menjadikan movinet tepat untuk aplikasi pengawasan *real-time* dan lingkungan yang terbatas sumber daya [1], [3].

Beberapa studi terbaru telah meneliti *deep learning* untuk pemantauan terkait sampah. Baru-baru ini, Alharbi et al. mengusulkan

sistem pengawasan berbasis *deep learning* yang mengintegrasikan YOLO dan MoViNet untuk deteksi perilaku membuang sampah di tempat umum, menunjukkan bahwa menggabungkan lokalisasi manusia dan pengenalan aksi video dapat mencapai deteksi pembuangan sampah yang akurat dan *real-time* dalam kondisi pengawasan sebenarnya. [1]. Sementara itu, Liu et al. memperkenalkan EcoDetect-YOLO, sebuah kerangka deteksi objek ringan untuk memantau limbah domestik di lingkungan yang kompleks, menyoroti efektivitas deteksi berbasis YOLO untuk aplikasi lingkungan [4]. Harada et al. juga mengembangkan algoritma berbasis YOLO yang ditingkatkan untuk deteksi sampah pada robot *mobile*, menunjukkan kinerja *real-time* yang efisien untuk pemantauan sampah di luar ruangan [5]. Meskipun studi-studi ini telah mencapai hasil yang menjanjikan, sebagian besar pendekatan yang ada terutama berfokus pada deteksi objek sampah atau menggunakan pengenalan aktivitas *end-to-end* tanpa secara eksplisit memisahkan lokalisasi manusia dari analisis aktivitas temporal.

Meskipun ada kemajuan, penelitian masih kurang dalam mengembangkan kerangka yang efektif yang menggabungkan deteksi manusia khusus dengan pengenalan aksi video untuk mendeteksi pembuangan sampah. Untuk mengenali perilaku membuang sampah, mendeteksi area manusia sebelum klasifikasi aktivitas dapat mengurangi informasi latar belakang yang tidak relevan dan memberikan fitur temporal yang lebih diskriminatif. Kerangka yang diusulkan menggunakan YOLOv11 secara eksplisit mengekstrak *Region of Interest* (ROI) manusia sebelum klasifikasi aktivitas menggunakan MoViNet-A0 berbeda dengan penelitian sebelumnya yang terutama berfokus pada deteksi objek atau pengenalan aktivitas *end-to-end*. *Pipeline* ini memungkinkan model pengenalan aksi untuk berkonsentrasi pada gerakan manusia sambil mengurangi gangguan latar belakang. Hal ini meningkatkan kemampuan untuk mengidentifikasi tindakan membuang sampah dalam rekaman CCTV.

Berdasarkan permasalahan ini, penelitian ini mengusulkan sistem deteksi aktivitas membuang sampah untuk video CCTV menggunakan MoViNet-A0 sebagai model pengenalan aksi video utama dengan deteksi manusia berbasis YOLOv11 untuk mengekstraksi *Region of Interest* (ROI) Manusia. YOLOv11 digunakan sebagai detektor manusia yang telah dilatih sebelumnya untuk melokalisasi individu, sementara MoViNet-A0 mengklasifikasikan urutan video yang diekstraksi menjadi aktivitas Membuang Sampah dan Normal. Efektivitas kerangka kerja yang diusulkan dievaluasi menggunakan *accuracy*, *precision*, *recall*, *F1-score*, dan *confusion matrix* untuk memvalidasi kemampuannya dalam mengenali aktivitas membuang sampah dari video CCTV pada kondisi sebenarnya.

## 2. TINJAUAN PUSTAKA

### 2.1 Deteksi Aktivitas Buang Sampah

Kegiatan membuang sampah sembarangan merupakan perilaku membuang sampah di tempat-tempat publik yang tidak sesuai, seperti taman, pinggir jalan, dan tepi sungai, daripada menggunakan fasilitas pembuangan sampah yang telah ditentukan. Perilaku ini menyebabkan pencemaran lingkungan, menurunkan tingkat kebersihan kota, dan mengancam keberlanjutan ekosistem [6]. Akibatnya, sistem pemantauan otomatis semakin penting untuk membantu manajemen lingkungan yang efektif.

Deteksi aktivitas membuang sampah berbeda dari deteksi objek sampah karena berfokus pada mengenali perilaku manusia selama proses pembuangan sampah dengan menganalisis berbagai bingkai video. Analisis berbasis video memberikan informasi kontekstual yang lebih kaya daripada deteksi gambar tunggal karena keputusan bergantung pada penampilan spasial dan gerakan temporal. Studi baru menunjukkan bahwa deteksi objek dan analisis video dapat membantu sistem pengawasan mengidentifikasi perilaku membuang sampah dengan lebih baik. Ini menunjukkan bahwa metode ini dapat digunakan untuk pemantauan lingkungan yang otomatis yang bergantung pada kamera CCTV [7], [8].

### 2.2 YOLOv11 untuk Deteksi Manusia

Deteksi objek merupakan tugas dasar dalam visi komputer yang bertujuan untuk mengidentifikasi lokasi dan kategori objek dalam gambar atau bingkai video. Algoritma *You Only Look Once* (YOLO) telah menjadi salah satu algoritma deteksi objek yang paling banyak diadopsi karena melakukan lokalisasi dan klasifikasi objek dalam arsitektur satu tahap dan memungkinkan kecepatan deteksi tinggi sambil mempertahankan akurasi yang kompetitif untuk aplikasi *real-time* [9]. Menurut Ali dan Zhang [9], perkembangan terbaru dari kerangka kerja YOLO, termasuk YOLOv11, telah meningkatkan kemampuan ekstraksi fitur, representasi fitur multi-skala, dan efisiensi komputasi, sehingga model ini cocok untuk sistem pengawasan cerdas. Demikian pula, Sapkota et al. [10] melaporkan bahwa YOLOv11 memperkenalkan penyempurnaan arsitektur yang meningkatkan keseimbangan antara kecepatan inferensi dan akurasi deteksi sambil mendukung penerapan pada perangkat edge dan platform komputasi berkinerja tinggi.

Dalam penelitian ini, YOLOv11 digunakan hanya sebagai detektor manusia untuk melokalisasi orang dalam bingkai video CCTV. Kotak pembatas manusia yang terdeteksi kemudian dipotong menjadi *Region of Interest* (ROI) sebelum diproses oleh model MoViNet-A0 untuk pengenalan aktivitas. Langkah pra-pemrosesan ini memungkinkan model pengenalan aktivitas untuk fokus pada gerakan manusia sambil meminimalkan gangguan latar belakang. Aldi et al. [11] menunjukkan bahwa YOLOv11 mencapai kinerja yang andal untuk deteksi dan penghitungan manusia secara *real-time* dalam video pengawasan yang menunjukkan bahwa model ini cocok sebagai modul lokalisasi manusia dalam sistem pemantauan cerdas. Selain itu, studi-studi terbaru telah menunjukkan bahwa YOLOv11 juga telah berhasil diadopsi dalam kerangka pengenalan aksi manusia dan menyoroti efektivitasnya sebagai ekstraktor fitur yang kuat untuk aplikasi visi berbasis pengawasan [12].

### 2.3 MoViNet-A0 untuk Pengenalan Aktivitas

*Human Action Recognition* (HAR) merupakan fungsi visi komputer yang menggunakan informasi spasial dan temporal

untuk secara otomatis mengenali tindakan manusia dari data video berurutan. Pengenalan aksi video berbeda dari klasifikasi gambar, yang bergantung pada satu gambar. Pengenalan aksi video menganalisis hubungan antara berbagai bingkai untuk mengidentifikasi pola gerakan yang menggambarkan tindakan manusia tertentu. Akibatnya, HAR telah banyak digunakan dalam sistem pengawasan otomatis karena memungkinkan pemahaman otomatis tentang perilaku manusia untuk berbagai tujuan, seperti pengawasan lingkungan, keselamatan publik, dan pemantauan lalu lintas [13], [14].

*Mobile Video Networks* (MoViNet) merupakan arsitektur yang dirancang khusus untuk meningkatkan pemahaman video sambil mempertahankan kemampuan pengenalan yang kompetitif. Mekanisme *stream buffering* dan *causal convolution* yang ditambahkan oleh MoViNet memungkinkan inferensi online pada video *streaming* dengan konsumsi memori yang jauh lebih rendah dibandingkan dengan jaringan *conventional 3D convolutional* [15].

Varian terkecil dari keluarga MoViNet yaitu MoViNet-A0 dapat menyeimbangkan dengan baik efisiensi komputasi dan akurasi pengenalan, sehingga cocok untuk aplikasi *real-time* dan perangkat yang memiliki sumber daya yang terbatas [15]. Oleh karena itu, urutan video manusia yang telah dipotong diklasifikasikan menjadi aktivitas Membuang Sampah dan Normal setelah *Region of Interest* (ROI) diekstraksi menggunakan YOLOv11 lalu menggunakan MoViNet-A0 sebagai model utama.

## 2.4 Metriks Evaluasi Kinerja

Evaluasi kinerja model dilakukan menggunakan *confusion matrix* yang menghasilkan nilai *accuracy*, *precision*, *recall*, dan *F1-score*. Keempat metrik tersebut dihitung berdasarkan nilai *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) [16], [17]. Metrik-metrik ini umumnya digunakan untuk mengevaluasi sistem deteksi dan klasifikasi video berbasis *deep learning* dengan mengukur kemampuan model dalam mengenali kelas target dengan benar [1].

*Accuracy* merupakan proporsi keseluruhan dari sampel yang diklasifikasikan dengan benar

di antara semua sampel pengujian, seperti yang didefinisikan dalam Persamaan (1).

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (1)$$

*Precision* merupakan metrik yang mengukur proporsi sampel positif yang diprediksi dengan benar dibandingkan dengan semua sampel yang diprediksi sebagai positif, sementara *Recall* mengukur kemampuan model untuk mendeteksi semua sampel positif yang sebenarnya. Presisi dan recall dihitung menggunakan Persamaan (2) dan Persamaan (3) secara berurutan.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

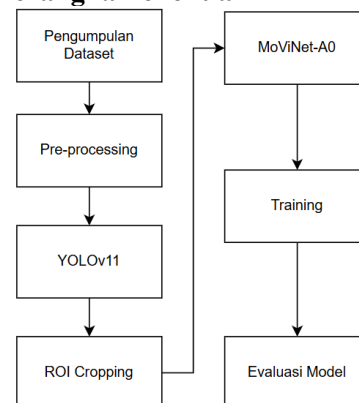
*F1-score* merupakan metrik yang digunakan untuk memberikan evaluasi yang seimbang antara *precision* dan *recall* dengan menggabungkan kedua metrik tersebut, seperti yang didefinisikan dalam Persamaan (4).

$$F1 - score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (4)$$

Dalam penelitian ini, metrik-metrik ini digunakan untuk mengevaluasi kinerja klasifier MoViNet-A0 dalam membedakan antara aktivitas buang sampah dan aktivitas normal dari urutan video CCTV. Selain itu, *confusion matrix* digunakan untuk menunjukkan hasil klasifikasi dengan menampilkan jumlah prediksi *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FN). Metode evaluasi ini menganalisis secara menyeluruh keandalan sistem deteksi manusia yang diusulkan berbasis YOLOv11 dan pengenalan aktivitas MoViNet-A0 [1], [18].

## 3. METODE PENELITIAN

### 3.1. Kerangka Penelitian



Gambar 1. Alur Kerangka Penelitian

Kerangka penelitian yang diusulkan pada penelitian ini terdiri dari beberapa tahapan yang saling berkaitan, dimulai dari pengumpulan dataset hingga evaluasi model. Alur penelitian ditunjukkan pada Gambar 1.

Tahap pertama adalah pengumpulan dataset, yaitu pengumpulan video CCTV yang berisi aktivitas Buang Sampah dan Normal. Dataset yang telah diperoleh kemudian memasuki tahap *pre-processing* untuk mempersiapkan data sebelum digunakan dalam proses pelatihan. Tahapan ini meliputi pembagian dataset menjadi data pelatihan, validasi, dan pengujian, penyesuaian format video, serta proses pelabelan sesuai dengan kelas aktivitas.

Selanjutnya, setiap *frame* video diproses menggunakan YOLOv11 untuk mendeteksi keberadaan manusia. Model ini hanya digunakan sebagai *human detector* sehingga menghasilkan *bounding box* yang menunjukkan lokasi objek manusia pada setiap *frame*. Berdasarkan hasil deteksi tersebut, sistem melakukan ROI (*Region of Interest*) *Cropping* dengan memotong area yang hanya berisi objek manusia. Proses ini bertujuan mengurangi pengaruh latar belakang sehingga informasi yang diterima model klasifikasi lebih berfokus pada gerakan manusia.



Gambar 2. *Person Crop* oleh YOLOv11

Video hasil *cropping* kemudian digunakan sebagai masukan bagi model MoViNet-A0. Pada tahap *training*, model mempelajari karakteristik gerakan dari setiap kelas aktivitas untuk membangun model klasifikasi yang mampu membedakan aktivitas Buang Sampah dan Normal. Setelah proses pelatihan selesai, performa model dievaluasi menggunakan data uji melalui metrik *accuracy*, *precision*, *recall*, *F1-score*, dan *confusion matrix* untuk mengetahui kemampuan model dalam mengenali aktivitas membuang sampah pada video CCTV.

### 3.2. Dataset

Dataset penelitian ini terdiri dari rekaman CCTV dari wilayah Sungai Sekanak, Palembang, Indonesia. Untuk mewakili lingkungan lokasi studi, video-video tersebut direkam pada siang hari di bawah kondisi pengawasan luar ruangan yang sebenarnya. Setiap video dianotasi ke dalam dua kategori aktivitas yaitu Buang Sampah dan Normal yang dikategorikan berdasarkan perilaku manusia yang diamati.

Dataset yang digunakan pada penelitian ini sebanyak 1.413 dataset video dengan total masing-masing kelas Buang Sampah sebanyak 548 dataset dan kelas Normal sebanyak 865 dataset. Distribusi dataset lebih lengkap dapat dilihat pada tabel 1 dibawah ini.

Tabel 1. Distribusi Dataset

| Kelas        | Train<br>ing | Validasi<br>on | Testi<br>ng | Total       |
|--------------|--------------|----------------|-------------|-------------|
| Buang Sampah | 384          | 82             | 82          | 548         |
| Normal       | 605          | 130            | 130         | 865         |
| <b>Total</b> | <b>989</b>   | <b>212</b>     | <b>212</b>  | <b>1413</b> |

Dataset dibagi menjadi tiga subset untuk memastikan pengembangan dan evaluasi model yang andal yaitu *training*, *validation*, dan *testing* dengan rasio 70:15:15. Set *training* mempelajari pola aktivitas, set *validation* mengoptimalkan parameter model selama *training*, dan set *testing* menilai kinerja generalisasi metode yang diusulkan pada data yang sebelumnya belum terlihat.



Gambar 3. Pengumpulan Dataset

### 3.3. Prapemrosesan Data

Pada tahap prapemrosesan data, seluruh video CCTV dipersiapkan agar sesuai dengan format masukan yang dibutuhkan oleh model MoViNet-A0. Proses ini bertujuan untuk menghasilkan data yang seragam sehingga proses pelatihan dapat berjalan secara optimal.

Tahapan prapemrosesan meliputi *video clipping*, *frame resizing*, deteksi manusia menggunakan YOLOv11, *ROI cropping*, serta pembentukan urutan frame (*sequence frame*).

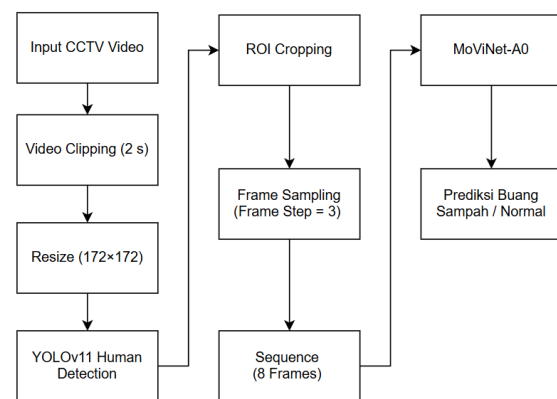
Video CCTV terlebih dahulu dipotong menjadi beberapa klip berdurasi 2 detik, di mana setiap klip hanya merepresentasikan satu aktivitas, yaitu Buang Sampah atau Normal. Selanjutnya, setiap *frame* pada klip video diubah ukurannya menjadi  $224 \times 224$  piksel agar sesuai dengan ukuran masukan model MoViNet-A0 sekaligus mengurangi kebutuhan komputasi selama proses pelatihan.

Setelah proses *resizing*, setiap *frame* diproses menggunakan YOLOv11 untuk mendeteksi objek manusia. Pada penelitian ini, YOLOv11 digunakan sebagai model pre-trained sehingga tidak dilakukan proses pelatihan ulang. Hasil deteksi berupa *bounding box* dimanfaatkan untuk melakukan *Region of Interest (ROI) cropping*, yaitu memotong area yang hanya berisi objek manusia. Dengan demikian, informasi latar belakang yang tidak berkaitan dengan aktivitas dapat diminimalkan sehingga model lebih berfokus pada gerakan manusia.

Tahap terakhir adalah pembentukan *sequence frame* sebagai masukan bagi MoViNet-A0. Sebanyak 16 *frame* dipilih secara merata (*uniform sampling*) dari setiap klip video sehingga mampu merepresentasikan keseluruhan aktivitas yang terjadi. Urutan *frame* tersebut kemudian digunakan sebagai input model MoViNet-A0 untuk mengekstraksi informasi spasial dan temporal dalam mengklasifikasikan aktivitas Buang Sampah dan Normal.

### 3.4. Metode yang Diusulkan

Metode yang diusulkan pada penelitian ini menggabungkan YOLOv11 sebagai pendeteksi manusia dengan MoViNet-A0 sebagai model utama untuk pengenalan aktivitas. Sebagai detailnya dapat dilihat pada Gambar 2 dibawah ini.



Gambar 4. Metode yang Diusulkan

Gambar 4 menunjukkan video CCTV yang telah melalui tahap preprocessing diproses menggunakan YOLOv11 untuk mendeteksi keberadaan manusia pada setiap frame. Hasil deteksi berupa *bounding box* kemudian digunakan untuk melakukan *Region of Interest (ROI) cropping* sehingga hanya area yang berisi objek manusia yang dipertahankan.

Selanjutnya, hasil *cropping* disusun menjadi urutan sebanyak 16 *frame* yang dipilih menggunakan metode *uniform sampling*. Teknik ini memastikan bahwa *frame* yang digunakan merepresentasikan keseluruhan durasi video secara merata. Urutan frame tersebut kemudian digunakan sebagai masukan bagi model MoViNet-A0 untuk mempelajari hubungan spasial dan temporal dalam membedakan aktivitas Buang Sampah dan Normal. Dengan memanfaatkan hasil *ROI cropping*, model dapat lebih berfokus pada pergerakan manusia sehingga pengaruh informasi latar belakang yang tidak relevan dapat diminimalkan.

### 3.5. Pelatihan Model

Model MoViNet-A0 dilatih menggunakan dataset hasil prapemrosesan untuk mempelajari karakteristik aktivitas Buang Sampah dan Normal. Selama proses pelatihan, beberapa *hyperparameter* ditentukan untuk mengoptimalkan proses pembelajaran model. Konfigurasi pelatihan yang digunakan pada penelitian ini ditunjukkan pada Tabel 2.

Tabel 2. Konfigurasi Pelatihan Model

| Parameter        | Value                   |
|------------------|-------------------------|
| Input Resolution | $224 \times 224$ pixels |
| Number of Frames | 16                      |
| Batch Size       | 4                       |

|                             |                                              |
|-----------------------------|----------------------------------------------|
| Head Training Epoch         | 15                                           |
| Fine-tuning Epoch           | 35                                           |
| Optimizer                   | AdamW                                        |
| Learning Rate (Head)        | 0.001                                        |
| Learning Rate (Fine-tuning) | 0.00001                                      |
| Loss Function               | Sparse Categorical Crossentropy              |
| Hardware                    | AMD Ryzen 5 5500U, Radeon Graphics, RAM 8 GB |

Tabel 2 menunjukkan konfigurasi pelatihan yang digunakan pada penelitian ini. Model MoViNet-A0 menerima masukan berupa urutan 16 *frame* berukuran  $224 \times 224$  piksel yang dipilih secara merata (*uniform sampling*) dari setiap klip video. Konfigurasi ini dipilih untuk mempertahankan informasi temporal yang cukup dalam merepresentasikan aktivitas sekaligus menjaga efisiensi komputasi.

Proses pelatihan dilakukan dalam dua tahap, yaitu *head training* selama 15 epoch dan *fine-tuning* selama 35 epoch. Tahap pertama bertujuan melatih lapisan klasifikasi yang baru ditambahkan, sedangkan tahap kedua digunakan untuk menyempurnakan bobot model dengan learning rate yang lebih kecil sehingga proses pembelajaran menjadi lebih stabil.

Optimizer Adam digunakan karena mampu memperbarui bobot secara adaptif sehingga mempercepat konvergensi model. Selain itu, *Sparse Categorical Crossentropy* dipilih sebagai fungsi *loss* untuk menghitung kesalahan prediksi pada proses klasifikasi dua kelas, sedangkan *accuracy* digunakan sebagai metrik utama selama proses pelatihan untuk memantau kinerja model.

### 3.6. Evaluasi Kinerja

Tahap evaluasi kinerja dilakukan untuk mengukur kemampuan model MoViNet-A0 dalam mengklasifikasikan aktivitas Buang Sampah dan Normal pada data uji. Evaluasi dilakukan setelah proses pelatihan selesai dengan membandingkan hasil prediksi model terhadap label sebenarnya. Pengukuran performa model menggunakan beberapa metrik

evaluasi, yaitu *accuracy*, *precision*, *recall*, *F1-score*, dan *confusion matrix*.

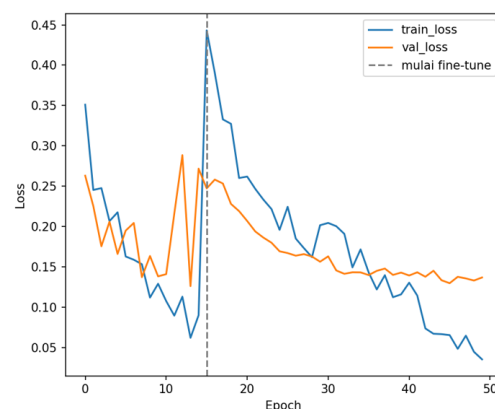
*Accuracy* digunakan untuk mengukur tingkat ketepatan model secara keseluruhan dalam mengklasifikasikan kedua kelas aktivitas. Namun, nilai *accuracy* saja belum cukup untuk menggambarkan performa model, terutama apabila distribusi data antar kelas tidak seimbang. Oleh karena itu, digunakan *precision* untuk menunjukkan ketepatan model dalam mengidentifikasi aktivitas Buang Sampah, sedangkan *recall* digunakan untuk mengukur kemampuan model dalam mendeteksi seluruh aktivitas Buang Sampah yang sebenarnya terjadi. Selanjutnya, *F1-score* digunakan sebagai metrik yang menyeimbangkan nilai *precision* dan *recall* sehingga memberikan gambaran performa model yang lebih komprehensif.

Selain itu, *confusion matrix* digunakan untuk menganalisis hasil klasifikasi secara lebih rinci dengan menunjukkan jumlah prediksi yang benar maupun salah pada masing-masing kelas. Melalui *confusion matrix*, dapat diketahui nilai *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) sehingga kesalahan klasifikasi yang dilakukan model dapat dianalisis secara lebih mendalam. Hasil evaluasi tersebut kemudian digunakan sebagai dasar untuk menilai efektivitas metode yang diusulkan dalam mendeteksi aktivitas membuang sampah pada video CCTV.

## 4. HASIL DAN PEMBAHASAN

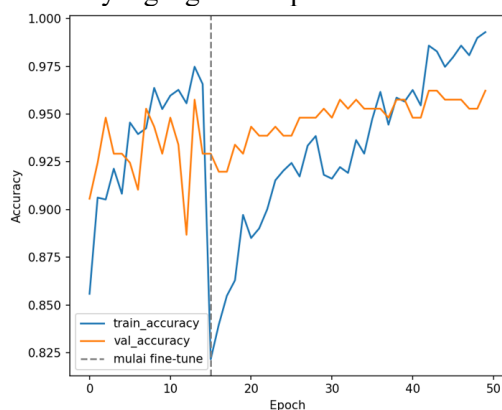
### 4.1. Hasil Pelatihan MoViNet-A0

Hasil pelatihan model MoViNet-A0 dengan menggunakan konfigurasi yang sudah disesuaikan dapat dilihat dari grafik pada Gambar 3 dan Gambar 4 dibawah ini.



Gambar 5. *Training dan Validation Loss Model MoViNet-A0*

Pada Gambar 5 menunjukkan perubahan nilai *training loss* dan *validation loss* selama proses pelatihan model MoViNet-A0. Pada tahap awal (*head training*), kedua kurva menunjukkan tren menurun, yang mengindikasikan bahwa model mulai mempelajari pola aktivitas dari data pelatihan. Memasuki tahap *fine-tuning* pada epoch ke-15, nilai *training loss* sempat meningkat karena proses pelatihan mulai memperbarui bobot pada *backbone* MoViNet-A0. Namun, seiring bertambahnya epoch, nilai *training loss* kembali menurun secara bertahap hingga mencapai nilai yang rendah pada akhir pelatihan. Sementara itu, *validation loss* juga menunjukkan kecenderungan menurun dan kemudian relatif stabil. Hal ini menunjukkan bahwa model mampu meningkatkan proses pembelajaran tanpa mengalami peningkatan kesalahan yang signifikan pada data validasi.



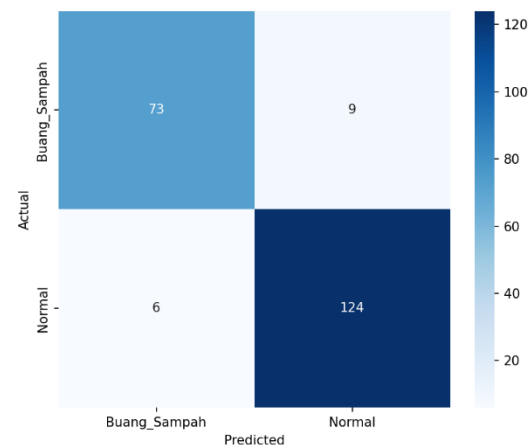
Gambar 6. *Training dan Validation Accuracy Model MoViNet-A0*

Pada Gambar 6 memperlihatkan perkembangan *training accuracy* dan *validation accuracy* selama proses pelatihan. Pada tahap *head training*, kedua kurva mengalami peningkatan yang menunjukkan bahwa model berhasil mempelajari karakteristik aktivitas Buang Sampah dan Normal. Setelah memasuki tahap *fine-tuning*, *training accuracy* meningkat secara bertahap hingga mencapai sekitar 99%, sedangkan *validation accuracy* berada pada kisaran 95–96% dan cenderung stabil hingga akhir pelatihan. Meskipun terdapat sedikit selisih antara kedua kurva, perbedaannya tidak terlalu besar sehingga menunjukkan bahwa model memiliki kemampuan generalisasi yang cukup baik terhadap data yang belum pernah dilihat sebelumnya.

Secara keseluruhan, tren penurunan nilai *loss* yang diikuti dengan peningkatan nilai *accuracy* menunjukkan bahwa proses pelatihan berlangsung dengan baik. Meskipun terdapat fluktuasi setelah tahap *fine-tuning*, model mampu mencapai kondisi yang stabil pada akhir pelatihan sehingga layak digunakan untuk proses evaluasi pada data uji.

## 4.2. Evaluasi Kinerja

Evaluasi kinerja model dilakukan setelah proses pelatihan selesai dengan membandingkan hasil prediksi model terhadap label sebenarnya. Pengukuran performa model menggunakan beberapa metrik evaluasi, yaitu *accuracy*, *precision*, *recall*, *F1-score*, dan *confusion matrix*.



Gambar 7. *Confusion Matrix*

Gambar 7 menunjukkan *confusion matrix* hasil klasifikasi model MoViNet-A0 terhadap dua kelas aktivitas, yaitu Buang Sampah dan Normal. Berdasarkan hasil tersebut, model memperoleh nilai TP (*True Positive*) sebanyak 73 video aktivitas Buang Sampah dengan benar dan TN (*True Negative*) sebanyak 124 video aktivitas Normal dengan benar. Di sisi lain, model memperoleh nilai FN (*False Negative*) sebanyak 9 video aktivitas Buang Sampah yang salah diprediksi sebagai Normal serta FP (*False Positive*) sebanyak 6 video aktivitas Normal yang salah diprediksi sebagai Buang Sampah. Jumlah kesalahan klasifikasi yang relatif sedikit menunjukkan bahwa model mampu membedakan kedua kelas dengan baik, meskipun masih terdapat beberapa aktivitas yang memiliki karakteristik gerakan serupa sehingga menyebabkan kesalahan prediksi.

Evaluasi kinerja model dapat dihitung menggunakan persamaan (1) sampai (4) seperti dibawah ini.

$$\begin{aligned} Accuracy &= \frac{TP+TN}{TP+TN+FN+FP} \\ &= \frac{73+124}{73+124+9+6} \\ &= \frac{197}{212} \\ &= 0.9292 \end{aligned} \quad (1)$$

$$\begin{aligned} Precision &= \frac{TP}{TP+FP} \\ &= \frac{73}{73+6} \\ &= \frac{73}{79} \\ &= 0.9241 \end{aligned} \quad (2)$$

$$\begin{aligned} Recall &= \frac{TP}{TP+FN} \\ &= \frac{73}{73+9} \\ &= \frac{73}{82} \\ &= 0.8902 \end{aligned} \quad (3)$$

$$\begin{aligned} F1 - score &= 2 \times \frac{Recall \times Precision}{Recall + Precision} \\ &= 2 \times \frac{0.8902 \times 0.9241}{0.8902 + 0.9241} \\ &= 2 \times \frac{0.8226}{1.8143} \\ &= 0.9068 \end{aligned} \quad (4)$$

Secara detail hasil keempat metrik ini disajikan dalam Tabel 3 dibawah ini.

Tabel 3. *Classification Report*

|                 | Precision | Recall | F1-score |
|-----------------|-----------|--------|----------|
| Buang Sampah    | 0.9241    | 0.8902 | 0.9068   |
| Normal          | 0.9323    | 0.9538 | 0.9430   |
| <b>Accuracy</b> | 0.9292    |        |          |

Berdasarkan Tabel 3, model memperoleh *accuracy* sebesar 92,92%, yang menunjukkan bahwa sebagian besar data uji berhasil diklasifikasikan dengan benar. Pada kelas Buang Sampah, model menghasilkan *precision* sebesar 92,41%, *recall* sebesar 89,02%, dan *F1-score* sebesar 90,68%. Nilai *precision* yang tinggi menunjukkan bahwa sebagian besar prediksi aktivitas buang sampah merupakan prediksi yang benar, sedangkan nilai *recall* yang sedikit lebih rendah mengindikasikan bahwa masih terdapat beberapa aktivitas Buang Sampah yang belum berhasil dikenali oleh model.

Sementara itu, pada kelas Normal, model memperoleh *precision* sebesar 93,23%, *recall*

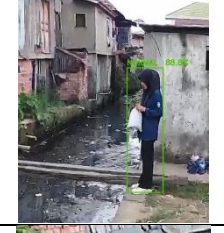
sebesar 95,38%, dan *F1-score* sebesar 94,30%, yang menunjukkan kemampuan model yang sangat baik dalam mengenali aktivitas normal. Secara keseluruhan, nilai *accuracy*, *precision*, *recall*, dan *F1-score* yang tinggi menunjukkan bahwa model MoViNet-A0 yang dikombinasikan dengan deteksi manusia berbasis YOLOv11 mampu memberikan performa yang baik dalam mendeteksi aktivitas buang sampah pada video CCTV.

#### 4.3. Hasil Deteksi Aktivitas Buang Sampah

Pengujian aktivitas tidak membuang sampah (Normal) dan membuang sampah dilakukan secara *real time* di siang hari berdiri. Hasil pengujian dilakukan secara langsung dengan menggunakan kamera CCTV agar mengetahui hasil deteksi yang tidak pernah dilihat sebelumnya. Selain itu, pengujian dimaksudkan untuk mengevaluasi konsistensi performa model terhadap variasi pencahayaan dan posisi objek. Dengan adanya pengujian ini, dapat diketahui sejauh mana sistem mampu mendeteksi aktivitas secara akurat dalam kondisi nyata yang berbeda-beda. Pada Tabel 4 merupakan hasil deteksi dari aktivitas tidak membuang sampah dan membuang sampah pada siang hari.

Tabel 4. Hasil Deteksi Pengujian

| No | Hasil Deteksi                                                                        | Prediksi     | Status |
|----|--------------------------------------------------------------------------------------|--------------|--------|
| 1. |  | Buang Sampah | Benar  |
| 2. |  | Buang Sampah | Benar  |
| 3. |  | Buang Sampah | Benar  |

|     |                                                                                     |              |       |
|-----|-------------------------------------------------------------------------------------|--------------|-------|
| 4.  |    | Buang Sampah | Benar |
| 5.  |    | Normal       | Benar |
| 6.  |    | Normal       | Benar |
| 7.  |   | Normal       | Benar |
| 8.  |  | Normal       | Salah |
| 9.  |  | Buang Sampah | Salah |
| 10. |  | Buang Sampah | Salah |

#### 4.4. Pembahasan

Berdasarkan Tabel 4, pengujian langsung menggunakan kamera CCTV pada kondisi siang hari terhadap 10 sampel aktivitas (4 buang

sampah, 3 normal murni, dan 3 kasus tambahan pada baris 8–10) menghasilkan 7 prediksi benar dan 3 prediksi salah, atau akurasi sebesar 70% pada pengujian lapangan ini. Angka ini lebih rendah dibandingkan akurasi hasil pelatihan model sebesar 92.9% yang menunjukkan adanya domain gap antara data pelatihan (video terkurasi) dan data pengujian *real-time* di lapangan.

Pada data ke-8, sistem menghasilkan prediksi Buang Sampah pada *bounding box* yang ditandai dengan kotak merah sementara aktivitas sebenarnya adalah normal, sehingga tercatat sebagai kesalahan *false positive*. Kesalahan ini kemungkinan besar dipicu oleh gestur objek yang menyerupai pola gerakan membuang sesuatu misalnya membungkuk atau mengayunkan tangan, padahal aktivitas tersebut bukan tindakan membuang sampah. Hal ini mengindikasikan bahwa model MoViNet-A0 masih rentan terhadap ambiguitas pola gerakan jangka pendek pada klip 16 *frame*, terutama saat gerakan tubuh objek tumpang tindih secara visual dengan pola buang sampah.

Pada data ke-9 dan ke-10, sistem memprediksi aktivitas sebagai normal padahal aktivitas sebenarnya adalah membuang sampah, sehingga tergolong kesalahan *false negative*. Kedua kasus ini menunjukkan pola visual serupa, yaitu objek berada pada posisi yang relatif jauh dari kamera dan sebagian tubuh tertutup vegetasi atau elemen bangunan di tepi sungai. Kondisi ini menurunkan kualitas informasi visual yang diterima MoViNet-A0 untuk membedakan pola gerakan membuang sampah dari gerakan tubuh normal, sehingga *confidence score* prediksi kemungkinan berada di bawah *threshold* deteksi yang telah ditetapkan pada sistem.

Penurunan akurasi dari 92.9% (data pelatihan) menjadi 70% (pengujian lapangan) mengonfirmasi bahwa performa model sangat dipengaruhi oleh kondisi lingkungan nyata yang lebih bervariasi dibanding data pelatihan, khususnya jarak objek terhadap kamera, oklusi oleh vegetasi/bangunan, dan ambiguitas pola gerakan jangka pendek. Temuan ini menjadi dasar rekomendasi untuk pengujian multi-hari lanjutan guna memperoleh sampel kegagalan yang lebih representatif, serta pertimbangan penyesuaian *threshold confidence* atau penambahan augmentasi data pelatihan dengan variasi jarak dan oklusi objek untuk

meningkatkan generalisasi model pada kondisi lapangan.

## 5. KESIMPULAN

Berdasarkan hasil pembahasan yang dilakukan, dapat disimpulkan ke dalam beberapa poin sebagai berikut.

- Sistem deteksi aktivitas membuang sampah berhasil dikembangkan dengan mengintegrasikan YOLOv11 sebagai pendeteksi manusia untuk mengekstraksi *Region of Interest* (ROI) dan MoViNet-A0 sebagai model pengenalan aktivitas yang mengklasifikasikan aktivitas menjadi Buang Sampah dan Normal.
- Hasil evaluasi pada data uji menunjukkan performa yang baik, dengan *accuracy* sebesar 92,92%. Pada kelas Buang Sampah, model memperoleh *precision* 92,41%, *recall* 89,02%, dan *F1-score* 90,68%, sedangkan pada kelas Normal diperoleh *precision* 93,23%, *recall* 95,38%, dan *F1-score* 94,30%. Hasil tersebut menunjukkan bahwa kombinasi YOLOv11 dan MoViNet-A0 mampu membedakan kedua aktivitas secara efektif.
- Pengujian secara *real-time* menggunakan kamera CCTV menghasilkan akurasi sebesar 70%, dengan kesalahan klasifikasi yang umumnya dipengaruhi oleh jarak objek terhadap kamera, oklusi akibat vegetasi atau elemen bangunan, serta kemiripan pola gerakan antara aktivitas Buang Sampah dan Normal. Hal ini menunjukkan bahwa validasi pada kondisi operasional nyata tetap diperlukan untuk mengevaluasi keandalan sistem di luar data uji.
- Penelitian selanjutnya disarankan untuk memperluas variasi data pelatihan, mencakup kondisi pencahayaan malam hari, variasi jarak dan tingkat oklusi objek, serta mengevaluasi penyesuaian *confidence threshold* secara adaptif guna meningkatkan performa sistem pada berbagai kondisi pengawasan CCTV.

## UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada orang tua penulis yang selalu mendukung penulis dalam setiap langkahnya, dan kepada Bapak Dr. RD. Kusumanto, S.T., M.M., dan

Ibu Dr. Nyayu Latifah Husni, S.T., M.T., serta seluruh pihak-pihak terkait yang telah memberikan dukungan terhadap penelitian ini.

## DAFTAR PUSTAKA

- [1] E. Alharbi, G. Alsulami, S. Aljohani, W. Alharbi, and S. Albaradei, "Real-time detection and monitoring of public littering behavior using deep learning for a sustainable environment," *Sci. Rep.*, vol. 15, no. 3000, pp. 1–11, 2025, doi: 10.1038/s41598-024-77118-x 1.
- [2] N. L. Husni *et al.*, "Pose Detection and Recurrent Neural Networks for Monitoring Littering Violations," *Eng.*, vol. 4, no. 4, pp. 2722–2740, 2023, doi: 10.3390/eng4040155.
- [3] Tarique Hussain, Zulfiqar Ali Memon, Rizwan Qureshi, and Tanvir Alam, "EMO-MoViNet: Enhancing Action Recognition in Videos with EvoNorm, Mish Activation, and Optimal Frame Selection for Efficient Mobile Deployment," *Sensors*, vol. 23, no. 19, p. 8106, 2023, doi: 10.3390/s23198106.
- [4] S. Liu, R. Chen, M. Ye, J. Luo, D. Yang, and M. Dai, "EcoDetect-YOLO: A Lightweight, High-Generalization Methodology for Real-Time Detection of Domestic Waste Exposure in Intricate Environmental Landscapes," *Sensors*, vol. 24, no. 14, p. 4666, 2024, doi: 10.3390/s24144666.
- [5] R. Harada, T. Oyama, K. Fujimoto, and T. Shimizu, "Trash Detection Algorithm Suitable for Mobile Robots Using Improved YOLO," *J. Adv. Comput. Intell. Intell. Informatics*, vol. 27, no. 4, pp. 622–631, 2023, doi: 10.20965/jaciii.2023.p0622.
- [6] Y. Z. Ulloa-torrealba, A. Schmitt, M. Wurm, A. Schmitt, and M. Wurm, "Litter on the streets - solid waste detection using VHR images," *Eur. J. Remote Sens.*, vol. 56, no. 1, p. 2176006, 2023, doi: 10.1080/22797254.2023.2176006.
- [7] U. Nurul, J. Paiton, U. N. Jadid, R. E. Pawening, and U. N. Jadid, "Pembuang Sampah Menggunakan Metode You Only Look Once (YOLO)," *TRILOGI J. Ilmu Teknol. Kesehatan, dan Hum.*, vol. 4, no. 2, pp. 98–105, 2023, doi: 10.33650/trilogi.v4i2.6676.
- [8] T. Gelar, S. Fitriani, and S. Rachmat, "A Systematic Literature Review of YOLO and IoT Applications in Smart Waste Management," *Green Intell. Syst. Appl.*, vol. 5, no. 2, pp. 123–139, 2025, doi: 10.53623/gisa.v5i2.706.
- [9] M. L. Ali and Z. Zhang, "The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in

- Object Detection,” *Computers*, vol. 13, no. 12, p. 336, 2024, doi: 10.3390/computers13120336.
- [10] R. Sapkota, M. F. Rizwan, and Q. Chetan, “YOLO advances to its genesis : a decadal and comprehensive review of the You Only Look Once ( YOLO ) series,” *Artif. Intell. Rev.*, vol. 58, no. 274, pp. 1–83, 2025, doi: 10.1007/s10462-025-11253-3.
- [11] F. Aldi, I. Nozomi, and S. Rahmawati, “Real-Time Human Detection and Counting System Based on YOLOv11 for Smart Campus Environment,” *J. Electr. Technol. UMY*, vol. 9, no. 2, pp. 64–73, 2025, doi: 10.18196/jet.v9i2.29281.
- [12] N. I. Nayeem, S. Mahbuba, S. I. Disha, R. H. Buiyan, S. Rahman, and J. Uddin, “A YOLOv11-Based Deep Learning Framework for Multi-Class Human Action Recognition,” *Comput. Mater. Contin.*, vol. 85, no. 1, pp. 1541–1557, 2025, doi: 10.32604/cmc.2025.065061.
- [13] T. Sultana, A. Alam, and Y. Lee, “Human Action Recognition: A Taxonomy-Based Survey, Updates, and Opportunities,” *Sensors*, vol. 23, no. 4, pp. 1–40, 2023, doi: 10.3390/s23042182.
- [14] M. Kulbacki *et al.*, “Intelligent Video Analytics for Human Action Recognition : The State of Knowledge,” *Sensors*, vol. 23, no. 9, pp. 1–31, 2023, doi: 10.3390/s23094258.
- [15] D. Kondratyuk *et al.*, “MoViNets: Mobile Video Networks for Efficient Video Recognition,” *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 16020–16030, 2021.
- [16] O. Rainio, J. Teuho, and R. Klén, “Evaluation metrics and statistical tests for machine learning,” *Sci. Rep.*, vol. 14, p. 6086, 2024, doi: 10.1038/s41598-024-56706-x.
- [17] G. M. Foody, “Challenges in the real world use of classification accuracy metrics : From recall and precision to the Matthews correlation coefficient,” *PLoS One*, vol. 18, no. 10, p. e0291908, 2023, doi: 10.1371/journal.pone.0291908.
- [18] M. M. Taye, “Theoretical Understanding of Convolutional Neural Network : Concepts, Architectures, Applications, Future Directions,” *Computation*, vol. 11, no. 3, pp. 1–23, 2023, doi: 10.3390/computation11030052.