



Peningkatan Akurasi Klasifikasi Tutupan Lahan Menggunakan Histogram Spesifikasi dan Random Forest Berbasis Citra Sentinel-2

Petricia Oktavia*, Muhamad Meki Frindo

Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia

Email: ^{1,*}dosen00638@unpam.ac.id, ²dosen00678@unpam.ac.id

Email Penulis Korespondensi: dosen00638@unpam.ac.id

Abstrak—Dinamika konversi tutupan lahan di Kabupaten Penajam Paser Utara (PPU) mengalami percepatan drastis pasca ditetapkan kawasan ini sebagai deliniasi wilayah Ibu Kota Nusantara (IKN). Pemantauan spasial menggunakan citra satelit optik Sentinel-2 menghadapi tantangan berat berupa gangguan radiometrik atmosfer tropis, kabut tipis (*haze*), serta tingginya kebingungan spektral (*spectral confusion*) antara kanopi hutan alami dan perkebunan kelapa sawit yang memiliki kemiripan nilai biomassa klorofil. Penelitian ini bertujuan untuk meningkatkan akurasi dan stabilitas pemetaan tutupan lahan 6 kelas (Hutan, Vegetasi Budidaya, Permukiman, Badan Air, Lahan Terbuka, dan Sawit) melalui integrasi pra-pemrosesan *Histogram Specification (HistSpec)*, rekayasa 9 fitur spektral, serta klasifikasi berbasis *Random Forest (RF)*. Kinerja RF dievaluasi secara komparatif terhadap enam algoritma *Machine Learning* lainnya (*Extra Trees*, *Decision Tree*, *Gradient Boosting*, *Logistic Regression*, *Naive Bayes*, dan *SGD Classifier*) pada 675 titik sampel (472 data latih dan 203 data uji independen). Hasil eksperimen membuktikan bahwa integrasi *Histogram Specification* dan *Random Forest* mencapai performa terbaik dengan *Overall Accuracy (OA)* sebesar 79,80% (meningkat dari 79,31% pada skenario citra RAW) serta koefisien Cohen's Kappa 0,7549. Sebaliknya, algoritma *Gradient Boosting* terbukti mengalami *overfitting* pada citra mentah (akurasi menurun dari 83,25% pada RAW menjadi 78,33% pada HistSpec). Dengan efisiensi komputasi yang tinggi (~123 ms) serta sifat *scale-invariant*, RF terbukti sebagai model paling baik dalam memproduksi peta tematik tutupan lahan bebas distorsi untuk mendukung perencanaan tata ruang dan kelestarian ekologis kawasan IKN.

Kata Kunci: Histogram Spesifikasi; Klasifikasi Tutupan Lahan; Penajam Paser Utara; Random Forest; Citra Sentinel-2

Abstract—Land cover conversion dynamics in Penajam Paser Utara (PPU) Regency have accelerated drastically following its designation as the core and buffer zone of the Nusantara Capital City (IKN). Optical remote sensing monitoring using Sentinel-2 imagery faces severe physical constraints from tropical atmospheric interference, haze, and pronounced spectral confusion between natural rainforest canopies and oil palm plantations exhibiting overlapping chlorophyll biomass reflectance. This study aims to improve the accuracy and robustness of 6-class land cover classification (Forest, Cultivated Vegetation, Settlement, Water Body, Bare Land, and Oil Palm) by integrating Histogram Specification (HistSpec) radiometric normalization, feature engineering of 9 spectral attributes, and Random Forest (RF) ensemble classification. The performance of RF was systematically benchmarked against six alternative Machine Learning algorithms (Extra Trees, Decision Tree, Gradient Boosting, Logistic Regression, Naive Bayes, and SGD Classifier) using 675 points (472 training and 203 independent testing points). Experimental findings demonstrate that Histogram Specification coupled with Random Forest attained stable performance, Overall Accuracy (OA) of 79.80% (an improvement from 78.33% on the RAW scenario) and a Cohen's Kappa coefficient of 0.7549. In contrast, Gradient Boosting exhibited severe overfitting on raw imagery (accuracy plummeted from 83.25% on RAW to 78.33% on HistSpec). Given its high computational efficiency (~123 ms) and scale-invariance, Random Forest proves to be the most dependable model for generating distortion-free thematic land cover maps to support sustainable spatial planning in the IKN region.

Keywords: Histogram Specification; Land Cover Classification; Penajam Paser Utara; Random Forest; Sentinel-2 Imagery

1. PENDAHULUAN

Kabupaten Penajam Paser Utara (PPU) di Provinsi Kalimantan Timur saat ini menjadi episentrum transformasi tata ruang, konversi lahan, dan pembangunan infrastruktur paling masif di Indonesia seiring dengan penetapan sebagian wilayahnya sebagai Kawasan Inti Pusat Pemerintahan (KIPP) dan kawasan penyangga Ibu Kota Nusantara (IKN) [1], [2]. Akselerasi pembangunan fisik berskala mega, pembukaan lahan (*land clearing*), perluasan jaringan transportasi terpadu, serta pergeseran mosaik lanskap telah mengubah bentang alam secara cepat dan drastis [1]. Di samping pembangunan kawasan perkotaan baru, bentang daratan PPU didominasi oleh benturan ekologis antara formasi hutan hujan tropis alami dataran rendah dan kawasan konservasi di bagian barat dengan ekspansi luas area perkebunan kelapa sawit (*Elaeis guineensis*) serta vegetasi budidaya monokultur di dataran rendah bagian tengah dan selatan. Kondisi dinamik ini menuntut ketersediaan sistem pemantauan tutupan lahan (*land cover monitoring*) yang teratur, berakurasi tinggi, terotomatisasi, dan mampu menyajikan data geospasial mutakhir untuk mitigasi deforestasi dan pengawasan tata ruang [1], [3], [4], [5].

Satelit Sentinel-2 *Multispectral Instrument (MSI)* yang dioperasikan oleh *European Space Agency (ESA)* di bawah program *Copernicus* menyediakan keunggulan signifikan dalam pemantauan sumber daya daratan ekuatorial [4]. Wahana ini membawa sensor dengan 13 saluran spektral, meliputi resolusi spasial tinggi 10 meter pada saluran tampak (*visible*) dan inframerah dekat (*Near-Infrared/NIR*), resolusi 20 meter pada saluran *red-edge* dan *Short-Wave Infrared (SWIR)*, serta siklus orbit ulang (*revisit time*) 5 hari dengan konstelasi ganda Sentinel-2A dan Sentinel-2B [4]. Karakteristik resolusi spasial dan temporal yang tinggi ini sangat ideal untuk mendeteksi perubahan tutupan lahan multi-skala [4], [6]. Namun demikian, pemanfaatan citra satelit optik di wilayah tropis basah seperti Kalimantan Timur menghadapi kendala fisik yang persisten, yaitu tingginya kelembapan udara, keberadaan tutupan awan kumululus, kabut tipis (*haze*), serta variasi hamburan aerosol atmosferik [4], [7]. Distorsi ini menyebabkan nilai pantulan spektral (*Bottom-of-atmosphere surface reflectance*) mengalami pergeseran kurva dan penurunan kontras visual antar piksel yang signifikan, terutama pada citra mentah multitemporal [6].



Tantangan kritis lainnya dalam pemetaan tutupan lahan di Kabupaten PPU adalah fenomena kebingungan spektral (*spectral confusion*) yang sangat tinggi antar kelas vegetasi [3], [6]. Area pohon hutan hujan lebat dan kanopi perkebunan kelapa sawit dewasa memantulkan respons spektral yang sangat mirip pada saluran tampak dan inframerah standar karena keduanya memiliki konsentrasi klorofil dan indeks biomassa daun yang melimpah [3], [8], [5]. Di sisi lain, pembukaan lahan untuk koridor IKN sering kali menghasilkan nilai pantulan tanah terbuka yang tumpang tindih dengan atap pemukiman atau perkerasan jalan pedesaan [1], [6]. Ketidakmampuan memisahkan kelas-kelas tersebut secara akurat memicu tingginya kesalahan klasifikasi (*misclassification error*), baik berupa kesalahan komisi maupun omisi [6], [9].

Dalam dekade terakhir, integrasi algoritma *Machine Learning* dengan data penginderaan jauh telah merevolusi metodologi klasifikasi tutupan lahan [4], [10], [11]. Di antara berbagai model yang berkembang, *Random Forest* (RF) menjadi salah satu pendekatan ansambel paling populer karena beroperasi menggunakan teknik *bagging* (*bootstrap aggregating*) dari banyak pohon keputusan (*Decision Trees*), tidak memerlukan asumsi distribusi parametrik normal, serta memiliki ketahanan alami terhadap *outlier* dan autokorelasi spasial [10], [11], [12], [13]. Meskipun demikian, performa algoritma pembelajaran mesin sangat bergantung pada representasi radiometrik data masukan [7], [10]. Pada citra mentah (*RAW*) yang memuat distorsi iluminasi dan derau lokal tidak serupa, algoritma yang sensitif terhadap gradien kesalahan seperti *Gradient Boosting* sangat rentan mengalami *overfitting* [10]. Model cenderung menghafal pola derau lokal alih-alih mempelajari batas karakteristik biofisik vegetasi yang sebenarnya, menghasilkan gangguan bintik-bintik acak (*salt-and-pepper noise*) pada peta tematik [9], [10].

Untuk menstabilkan disparitas radiometrik tersebut, teknik penyesuaian distribusi nilai piksel melalui metode *Histogram Specification* (atau *Histogram Matching*) menawarkan solusi matematika yang terbukti kuat [7]. Berbeda dengan *histogram equalization* biasa yang kerap menimbulkan saturasi kontras ekstrem, *Histogram Specification* mentransformasikan fungsi distribusi kumulatif (*Cumulative Distribution Function/CDF*) citra target agar mereplikasi profil kurva distribusi radiometrik dari citra referensi yang bersih dan bebas awan [7]. Transformasi ini menyelaraskan rentang dinamis citra lintas waktu tanpa mengaburkan batas spasial objek. Di samping itu, peningkatan separabilitas spektral dapat diperkuat melalui rekayasa fitur (*feature engineering*) indeks komposit biofisik seperti NDVI, NDWI, SAVI, EVI, dan GNDVI [3], [14].

Walaupun keunggulan teoretis dari masing-masing metode tersebut telah banyak diteliti, studi komparatif komprehensif yang menguji keandalan *Random Forest* secara langsung terhadap beragam paradigma *Machine Learning* lainnya (termasuk *Extra Trees*, *Decision Tree*, *Gradient Boosting*, *Logistic Regression*, *Gaussian Naive Bayes*, dan *SGD Classifier*) pada skenario citra sebelum dan sesudah *Histogram Specification* untuk 6 kelas tutupan lahan wilayah penyangga IKN masih belum pernah dilaporkan secara terintegrasi [7], [10], [11], [15]. Secara spesifik, penelitian ini memberikan empat kontribusi ilmiah utama yang memperkaya kajian penginderaan jauh dan pembelajaran mesin terapan, yaitu (1) mengembangkan kerangka integrasi *Histogram Specification* (*HistSpec*) untuk normalisasi radiometrik citra Sentinel-2 pada wilayah tropis ekuatorial yang basah. Metode ini mampu menyelaraskan distribusi CDF, mengurangi pengaruh *haze*, dan menjaga konsistensi kontras reflektansi antarwaktu tanpa mengurangi resolusi spasial 10 meter. (2) menunjukkan adanya kecenderungan *overfitting* pada *Gradient Boosting* akibat sensitivitas terhadap *noise* atmosferik pada citra RAW, sekaligus membuktikan ketahanan *Random Forest* terhadap variasi skala dan derau. Perbandingan dilakukan dengan enam algoritma lain pada 675 titik *ground truth* independen. (3) menemukan pentingnya saluran *Near-Infrared* (NIR/B8) dan *Enhanced Vegetation Index* (EVI) dalam mengurangi *spectral confusion* antara hutan hujan alami dan perkebunan kelapa sawit monokultur di Kalimantan Timur. (4) menghasilkan peta tutupan lahan enam kelas Kabupaten Penajam Paser Utara beresolusi 10 meter dengan *noise salt-and-pepper* yang minimal. Hasil ini dapat menjadi baseline spasial untuk memantau perubahan tutupan lahan serta mendukung perencanaan tata ruang berkelanjutan di kawasan inti dan penyangga IKN.

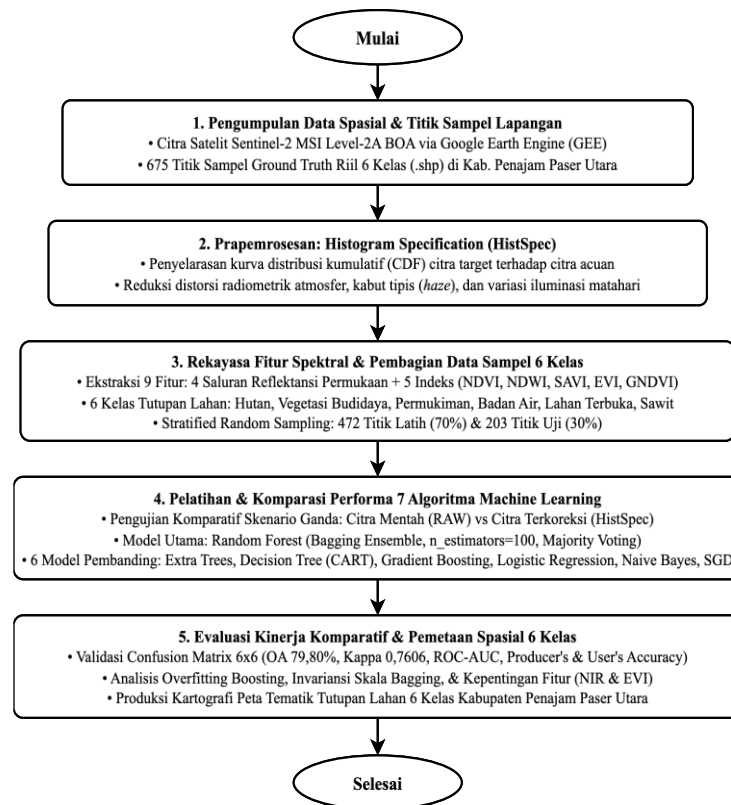
2. METODOLOGI PENELITIAN

Metodologi dalam penelitian ini dirancang secara terstruktur dan terstandarisasi untuk memastikan validitas ilmiah serta replikabilitas eksperimen geospasial. Rangkaian tahapan penelitian mencakup akuisisi data spasial Sentinel-2, normalisasi radiometrik *Histogram Specification*, rekayasa fitur spektral multi-indeks, penyusunan data sampel uji-latih 6 kelas, pelatihan dan komparasi 7 model *Machine Learning*, hingga validasi akurasi menggunakan matriks kesalahan dan analisis spasial kartografi.

2.1 Tahapan Penelitian

Alur metodologi penelitian diuraikan secara sistematis sebagaimana disajikan dalam diagram alir pada Gambar 1. Tahapan utama meliputi: (1) Pengumpulan Data Spasial: Pengunduhan citra Sentinel-2 MSI Level-2A *Bottom-of-Atmosphere* (BOA) melalui platform komputasi awan *Google Earth Engine* (GEE) [14] pada musim kering dengan tutupan awan minimal (< 10%) yang melingkupi wilayah administratif Kabupaten Penajam Paser Utara, serta penyiapan *shapefile* titik sampel, (2) Prapemrosesan *Histogram Specification* dengan ekstraksi kurva CDF citra target dan penyesuaian distribusinya terhadap citra acuan bebas awan untuk mereduksi distorsi kabut atmosferik dan variasi pencahayaan, (3) Ekstraksi 9 dimensi fitur yang terdiri dari 6 saluran reflektansi permukaan (B2, B3, B4, B8) dan 5 indeks komposit biofisik (NDVI, NDWI, SAVI, EVI, GNDVI) [4], [9], (4): Pengelompokan 675 titik darat ke dalam 6 kelas tutupan lahan (Hutan, Vegetasi Budidaya, Permukiman, Badan Air, Lahan Terbuka, dan Sawit) dengan pembagian

stratified split 70% data latih (472 titik) dan 30% data uji (203 titik) secara spasial independen [11]; (5) Pengujian komparatif antara *Random Forest* [12], [13] dan enam algoritma pembandingan (*Extra Trees*, *Decision Tree*, *Gradient Boosting*, *Logistic Regression*, *Gaussian Naive Bayes*, *SGD Classifier*) pada skenario citra RAW dan citra terkoreksi *HistSpec* [10], [11], [15]; serta (6) Validasi statistik menggunakan *Overall Accuracy*, matriks kesalahan (*Confusion Matrix*), koefisien Cohen's Kappa [9], analisis kurva ROC-AUC, evaluasi *feature importance*, serta pembuatan peta tematik tutupan lahan akhir.



Gambar 1. Diagram alir tahapan penelitian

2.2 Area Studi dan Karakteristik Data Spektral Sentinel-2

Wilayah penelitian berlokasi di Kabupaten Penajam Paser Utara, Provinsi Kalimantan Timur, yang terletak secara geografis pada koordinat 0°54'01" LS hingga 1°30'05" LS dan 116°19'30" BT hingga 116°56'35" BT [1]. Wilayah ini berbatasan dengan Selat Makassar dan Kota Balikpapan di timur, Kabupaten Kutai Kartanegara di utara dan barat, serta Kabupaten Paser di selatan. Data spektral yang dimanfaatkan adalah citra Sentinel-2 MSI Level-2A dengan koreksi atmosferik terintegrasi Sen2Cor [2]. Saluran spektral beresolusi spasial 10 meter diantaranya B2 Biru, B3 Hijau, B4 Merah, B8 NIR.

2.3 Metode Histogram Spesifikasi

Metode *Histogram Specification* mentransformasikan nilai derajat keabuan citra target $f(x,y)$ agar mengikuti distribusi probabilitas kumulatif citra acuan $g(x,y)$ [7]. Misalkan r menyatakan tingkat intensitas citra sumber dan z menyatakan tingkat intensitas citra acuan dalam rentang kontinu $[0, 1]$. Fungsi distribusi kumulatif (*Cumulative Distribution Function/CDF*) citra sumber $T(r)$ didefinisikan sebagai integral fungsi probabilitas densitasnya $p_r(w)$ [7]:

$$s = T(r) = \int p_r(w)dw, 0 \leq w \leq r \quad (1)$$

Secara analogis, fungsi distribusi kumulatif citra referensi yang ditargetkan $G(z)$ diformulasikan sebagai:

$$v = G(z) = \int p_z(t)dt, 0 \leq t \leq z \quad (2)$$

Dengan menetapkan kondisi ekuivalensi $s = v$, nilai intensitas terkoreksi z diperoleh melalui inversi transformasi distribusi kumulatif citra acuan terhadap nilai transformasi citra sumber [7]

$$z = G^{-1}(s) = G^{-1}[T(r)] \quad (3)$$

Pada domain digital diskrit, operasi ini diwujudkan melalui tabel pemetaan frekuensi kumulatif terakumulasi untuk setiap saluran spektral secara mandiri. Melalui normalisasi ini, deviasi radiometrik akibat kabut aerosol dan perbedaan elevasi matahari lintas tanggal perekaman dapat distandarkan tanpa merusak integritas batas spasial fitur permukaan.



2.4 Rekayasa Fitur Spektral dan Pemodelan Machine Learning

Ruang fitur yang digunakan dalam pemodelan mencakup 9 dimensi atribut spektral yang menggabungkan 4 saluran spektral reflektansi dan 5 indeks vegetasi, kelembapan, serta air [3], [8]. Algoritma Random Forest (RF) diimplementasikan sebagai pengklasifikasi utama berbasis ansambel bagging yang membangun B pohon keputusan independen [12]. Setiap pohon keputusan T_b dilatih menggunakan sampel bootstrap acak dari himpunan data latih [10], [12]. Pada setiap simpul percabangan (*node split*), subset acak fitur $m = \sqrt{p}$ dipilih (dengan $p = 11$) untuk meminimalkan korelasi antar-pohon dan menekan variansi model [11], [12].

$$C_R F(x) = \operatorname{argmax}_k \sum I(T_b(x) = k), k \in 1, \dots, K \quad (4)$$

di mana $K = 6$ menyatakan jumlah kelas tutupan lahan, dan $I(\cdot)$ adalah fungsi indikator biner. *Hyperparameter Random Forest* dikonfigurasi dengan $n_estimators = 100$, $max_depth = None$, kriteria *Gini Impurity*, dan bobot kelas seimbang ($class_weight = 'balanced'$). Untuk menguji ketahanan model, dibangun kerangka komparatif secara bersamaan terhadap 6 algoritma *Machine Learning* lainnya yakni *Extra Trees (Extremely Randomized Trees)* [16], *Decision Tree (CART)*, *Gradient Boosting (GBDT)*, *Logistic Regression (Multinomial Softmax)*, *Gaussian Naive Bayes*, dan *Stochastic Gradient Descent (SGD Classifier)*.

Data *ground truth* disusun sebanyak 675 titik sampel riil yang diekstraksi dari *shapefile* validasi terverifikasi lapangan dan interpretasi citra resolusi sangat tinggi Google Earth [14]. Data ini terbagi secara proporsional ke dalam 6 kelas tutupan lahan dengan rasio 70% data latih (472 titik) dan 30% data uji (203 titik), sebagaimana dirinci pada Tabel 1.

Tabel 1. Distribusi Data Sampel Titik Uji dan Latih 6 Kelas Tutupan Lahan Kabupaten PPU

Kode	Nama Kelas Tutupan Lahan	Deskripsi Fisik Objek	Sampel Latih (70%)	Sampel Uji (30%)	Total Titik
C1	Hutan	Kanopi hutan primer dan sekunder lebat, tutupan rapat	96	41	137
C2	Vegetasi Budidaya	Lahan pertanian pangan, hortikultura, dan semak vegetatif	60	25	85
C3	Permukiman	Kawasan perumahan, infrastruktur gedung, jalan, dan fasilitas umum	103	44	147
C4	Badan Air	Sungai, waduk, embung penampungan, teluk, dan estuari pesisir	60	25	85
C5	Lahan Terbuka	Tanah terbuka gundul, bukaan proyek IKN, dan tegalan pasca-panen	54	23	77
C6	Sawit	Hamparan monokultur perkebunan kelapa sawit produktif	101	43	144
-	Total Sampel	-	472	203	675

2.5 Validasi dan Evaluasi Kinerja Klasifikasi

Evaluasi kinerja klasifikasi diukur menggunakan matriks kesalahan (*Confusion Matrix*) berdimensi 6×6 terhadap 203 titik sampel uji independen [17]. Metrik evaluasi kuantitatif utama meliputi *Overall Accuracy* (OA) dan Koefisien Cohen's Kappa (κ) [9], [18]:

$$OA = (\sum x_{ii} / N) * 100 \quad (5)$$

$$\kappa = [N \sum x_{ii} - \sum (x_{i+} \cdot x_{+i})] / [N^2 - \sum (x_{i+} \cdot x_{+i})] \quad (6)$$

di mana x_{ii} adalah jumlah piksel kelas ke- i yang diklasifikasikan secara tepat, N adalah total titik uji ($N = 203$), x_{i+} adalah total marjinal baris (sampel referensi aktual), dan x_{+i} adalah total marjinal kolom (sampel prediksi model) [9], [18]. *Producer's Accuracy* (PA) dan *User's Accuracy* (UA) dihitung untuk masing-masing kelas guna mendiagnosis kesalahan omisi dan komisi, dilengkapi analisis kurva ROC-AUC untuk memvalidasi separabilitas probabilitas model [9], [11].

3. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil transformasi radiometrik *Histogram Specification*, analisis kepentingan fitur spektral, evaluasi komparatif mendalam antara *Random Forest* dan enam algoritma *Machine Learning* lainnya, matriks kesalahan 6 kelas, serta visualisasi kartografis tutupan lahan Kabupaten Penajam Paser Utara.

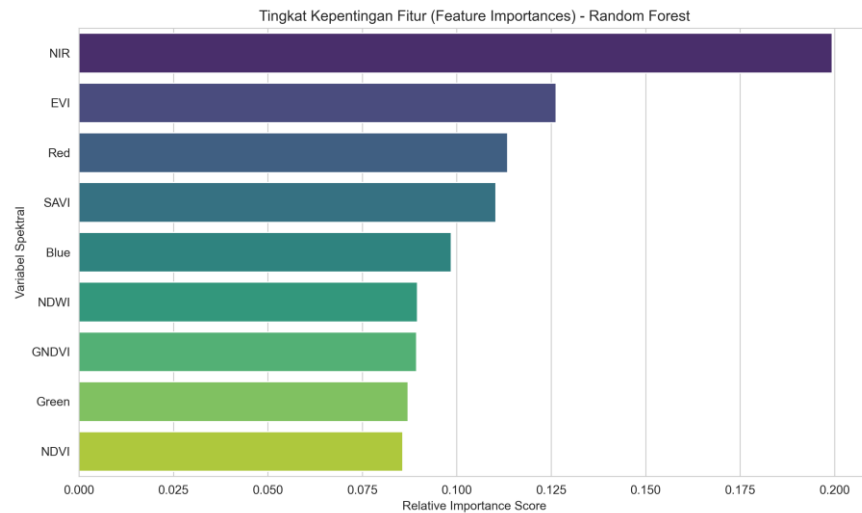
3.1 Hasil Transformasi Radiometrik Histogram Spesifikasi

Penerapan algoritma *Histogram Specification* berhasil menyelaraskan nilai reflektansi citra Sentinel-2 target terhadap citra acuan. Pada skenario citra mentah (RAW), citra Sentinel-2 wilayah PPU memperlihatkan dominasi sebaran atmosferik dan kabut tipis (*haze*) pada saluran tampak B2 (Biru) dan B3 (Hijau), yang ditandai oleh pergeseran nilai

minimum piksel ke arah intensitas yang lebih tinggi serta kurva histogram yang sempit dan asimetris. Kondisi ini menyebabkan kontras spektral antara objek kanopi vegetasi dan permukaan daratan non-vegetasi menjadi tereduksi [6], [7].

Setelah dilakukan transformasi *Histogram Specification*, rentang dinamis nilai digital melebar secara merata. Pada saluran inframerah dekat (B8), histogram citra terkoreksi mampu memisahkan puncak pantulan kanopi daun berkerapatan tinggi secara tegas dari spektrum tanah terbuka dan badan air. Reduksi kabut atmosferik dan perataan kontras global terbukti mengurangi ilusi variasi iluminasi tanpa mengubah nilai batas tepi (*edge preservation*) objek geospasial. Hal ini menciptakan ruang fitur yang jauh lebih konsisten bagi algoritma *machine learning* untuk menarik bidang pemisah (*decision boundary*) antar kelas tutupan lahan [10], [11].

Algoritma RF menyediakan fungsi untuk mengukur kontribusi relatif fitur masukan berbasis penurunan *Mean Decrease in Impurity/MDI* [11], [12]. Kontribusi masing-masing dari 9 fitur spektral divisualisasikan pada Gambar 2.



Gambar 2. Distribusi nilai fitur spektral pada pemodelan RF

Berdasarkan grafik pada Gambar 2, saluran inframerah dekat (B8 / NIR) menempati peringkat tertinggi dengan skor kepentingan sebesar 0,198, disusul secara ketat oleh indeks *Enhanced Vegetation Index* (EVI) dengan skor 0,126. Tingginya kontribusi B8 dan EVI membuktikan peranan penting dari radiasi inframerah dalam menembus kanopi dan merespons sebaran ganda pada struktur sel mesofil daun [3], [8], [19] [20]. Keunggulan EVI di atas NDVI membuktikan literatur penginderaan jauh wilayah tropis [19], formula EVI yang mengintegrasikan koefisien saluran biru mampu meredam gangguan aerosol atmosferik serta mencegah kejenuhan (*saturation effect*) yang kerap dialami NDVI pada hutan hujan tropis lebat dan perkebunan kelapa sawit berbiomassa tinggi [21].

3.2 Evaluasi Kinerja Komparatif Random Forest terhadap 7 Algoritma Machine Learning

Untuk memberikan justifikasi ilmiah yang kuat atas penunjukan *Random Forest* sebagai model definitif, kami melakukan pengujian komparatif menyeluruh terhadap tujuh algoritma *Machine Learning* yang mencakup empat paradigma pemodelan utama, yakni (1) ansambel berbasis pohon (*Random Forest*, *Extra Trees*, *Gradient Boosting*); (2) pohon keputusan tunggal (*Decision Tree/CART*); (3) model parametrik linier & jarak (*Logistic Regression*, *SGD Classifier*); serta (4) probabilistik independen (*Gaussian Naive Bayes*). Seluruh model diuji pada dua skenario data identik dengan menggunakan citra mentah (RAW) dan citra terkoreksi (HistSpec) menggunakan 203 titik sampel uji independen. Hasil perbandingan metrik kinerja dirangkum pada Tabel 2.

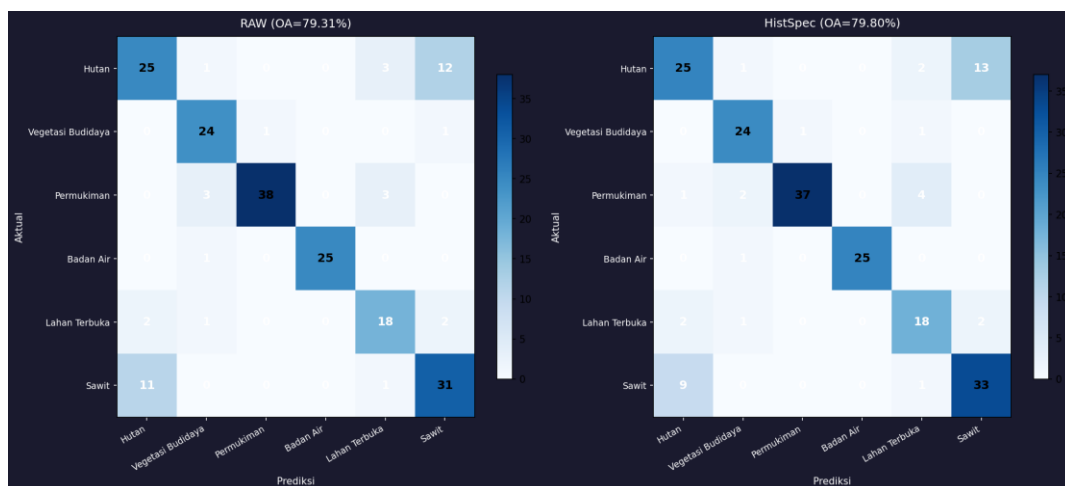
Tabel 2. Komparasi Kinerja Klasifikasi 7 Algoritma Machine Learning pada Skenario Citra RAW dan Citra Terkoreksi (HistSpec)

No	Algoritma Machine Learning	OA Skenario RAW (%)	OA Skenario HistSpec (%)	Perubahan (Δ OA)	Macro F1 HistSpec (%)	Waktu Latih (ms)
1	Random Forest (RF)	79,31%	79,80%	+1,47% (Meningkat & Stabil)	81,32%	123 ms
2	Extra Trees (ET)	78,33%	79,31%	+0,98% (Meningkat)	80,53%	76 ms
3	Gradient Boosting (GBDT)	83,25%	78,33%	-4,92% (Overfitting pada RAW)	78,98%	836 ms
4	Decision Tree (CART)	78,82%	72,91%	-5,91% (Sensitif Varians)	73,59%	4 ms
5	Logistic Regression	68,97%	68,47%	-0,50% (Rigiditas Linier)	69,74%	6 ms

No	Algoritma Machine Learning	OA Skenario RAW (%)	OA Skenario HistSpec (%)	Perubahan (Δ OA)	Macro F1 HistSpec (%)	Waktu Latih (ms)
6	SGD Classifier	56,65%	66,01%	+9,36% (Sensitif Skala)	66,84%	5 ms
7	Gaussian Naive Bayes	67,00%	66,50%	-0,50% (Kolinearitas Spektral)	66,96%	1 ms

Data pada Tabel 3 mengungkap sejumlah temuan yang sangat penting dalam domain klasifikasi penginderaan jauh multi-spektral:

- Random Forest membuktikan keunggulannya sebagai model terbaik pasca-HistSpec dengan mencapai *Overall Accuracy* 79,80% (162 dari 203 titik terklasifikasi benar), Macro F1 sebesar 81,32%, dan koefisien Cohen's Kappa 0,7549 [19]. Kinerja ini meningkat konsisten sebesar +1,47% dari skenario RAW (79,31%). Algoritma *Extra Trees* yang memiliki arsitektur serupa dengan pemilihan ambang batas acak murni juga mencatat peningkatan performa (dari 78,33% menjadi 79,31%, Macro F1 80,53%). Sifat *scale-invariant* pada pohon keputusan membuat model ini kebal terhadap pergeseran rentang skala reflektansi dan mampu mengekstraksi manfaat maksimal dari peningkatan separabilitas spektral hasil normalisasi histogram [10], [12].
- Pada citra mentah (RAW), *Gradient Boosting* menghasilkan akurasi semu yang tampak sangat tinggi (83,25%). Namun, ketika diuji pada citra terkoreksi HistSpec, akurasinya turun hingga sebesar -4,92% menjadi 78,33% (Macro F1 78,98%). Penurunan tajam ini merupakan bukti nyata terjadinya fenomena *overfitting*. Arsitektur sekuensial boosting yang secara berulang meminimalkan residu kesalahan (*residual loss*) menyebabkan pohon-pohon cabang tahap lanjut menghafal pola yang ada, distorsi kabut, dan anomali pada citra RAW tidak mampu mempelajari batas biofisik penutup lahan yang benar. Saat citra dibersihkan oleh *Histogram Specification*, gangguan tersebut terhapus, sehingga performa *boosting* menurun ke batas generalisasi alaminya. Sebaliknya, *Random Forest* terbukti tahan terhadap hal ini berkat mekanisme *bootstrap* agregasi dan *random feature subspace* [11], [12].
- Algoritma linier yang bergantung pada perhitungan gradien jarak (*SGD Classifier*) mendemonstrasikan tidak mampuan pada citra RAW yang menurun (56,65%), namun meningkat tinggi (+9,36%) hingga mencapai 66,01% pasca-spesifikasi histogram, membuktikan sensitivitasnya yang baik terhadap skala rentang data. Model *Logistic Regression* terhambat pada batas akurasi 68,47% karena pada bidang batas linier yang sulit diterapkan pada ruang spektral non-linier [10]. Hasil pada *Gaussian Naive Bayes* rendah di angka 66,50% akibat kesulitan menebak saluran spektral yang diganggu oleh autokorelasi dan multikolinearitas antar-saluran satelit (B3, B4, B8) dan indeks spektral turunannya.
- Dari segi efisiensi komputasi, RF hanya memerlukan waktu pelatihan 123 ms, hampir 7 kali lebih cepat dibandingkan *Gradient Boosting* yang menelan durasi hingga 836 ms. Untuk pemetaan kartografi regional skala luas seperti Kabupaten PPU yang memuat miliaran titik piksel spasial resolusi 10 meter, disparitas kecepatan ini menjadi faktor pendukung di mana arsitektur boosting menjadi tidak praktis (*computationally prohibitive*).



Gambar 3. Matriks konfusi hasil klasifikasi tutupan lahan 6 kelas model Random Forest pada skenario citra HistSpec

Tabel 3. Matriks Kesalahan (*Confusion Matrix* 6x6) Random Forest pada Citra Terkoreksi (HistSpec) 203 Titik Sampel Uji

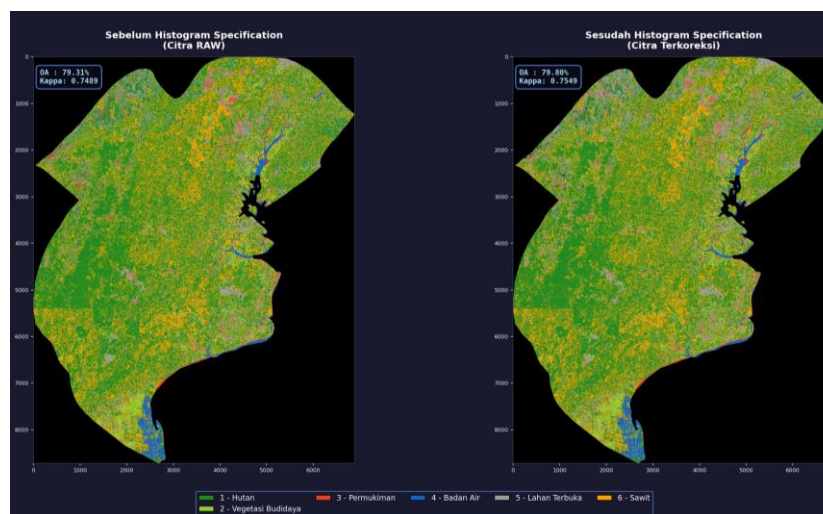
Kelas Aktual	Hutan	Veg. Budidaya	Permukiman	Badan Air	Lahan Terbuka	Sawit	Total Aktual	Producer's Acc (PA)
Hutan	25	1	0	0	2	13	41	60,98%
Vegetasi Budidaya	0	24	1	0	1	0	26	92,31%

Kelas Aktual	Hutan	Veg. Budidaya	Permukiman	Badan Air	Lahan Terbuka	Sawit	Total Aktual	Producer's Acc (PA)
Permukiman	1	2	37	0	4	0	44	84,09%
Badan Air	0	1	0	25	0	0	26	96,15%
Lahan Terbuka	2	1	0	0	18	2	23	78,26%
Sawit	9	0	0	0	1	33	43	76,74%
Total Prediksi	37	29	38	25	26	48	203	-
User's Acc (UA)	67,57%	82,76%	97,37%	100,0%	69,23%	68,75%	-	OA = 79,80%

Analisis terhadap matriks kesalahan pada Tabel 3 dan Gambar 3 memperlihatkan bahwa kelas Badan Air mencatatkan tingkat akurasi sempurna dengan *User's Accuracy* 100,0% dan *Producer's Accuracy* 96,15% (25 dari 26 titik terklasifikasi tepat). Hal ini dimungkinkan oleh penyerapan spektral total gelombang inframerah (B8) oleh molekul air serta kekhasan rentang indeks NDWI yang sangat tajam memisahkan air dari permukaan daratan [6], [8]. Kelas Permukiman juga meraih keandalan sangat tinggi dengan UA 97,37% dan PA 84,09% berkat saluran indeks SAVI yang efektif membedakan material atap seng dan beton dari tanah terbuka [19].

Pusat kebingungan spektral (*spectral confusion*) utama terlihat di zona interaksi antara kelas Hutan dan Sawit [5]. Sebanyak 13 titik sampel Hutan terprediksi sebagai Sawit (*omission error* kelas hutan), dan 9 titik sampel Sawit terprediksi sebagai Hutan (*commission error* kelas hutan). Fenomena ini secara ekologis dapat dipahami karena tajuk pohon kelapa sawit dewasa yang telah berumur di atas 8 tahun membentuk kanopi yang sangat rimbun dengan kandungan klorofil dan indeks biomassa yang menyerupai karakteristik spektral hutan sekunder dataran rendah [1], [3]. Kendati demikian, penerapan *Histogram Specification* terbukti menurunkan kesalahan klasifikasi ini, meningkatkan tebakan benar (*True Positive*) kelas Sawit dari 31 titik pada RAW menjadi 33 titik pada HistSpec (PA 76,74% dan UA 68,75%).

Peta tematik spasial tutupan lahan Kabupaten Penajam Paser Utara hasil proyeksi raster model *Random Forest* divisualisasikan secara komparatif pada Gambar 4.



Gambar 4. Perbandingan spasial peta tutupan lahan 6 kelas Kabupaten Penajam Paser Utara antara skenario citra RAW dan citra HistSpec

Peta tematik pada Gambar 4 menunjukkan bahwa pada skenario citra RAW (sebelah kiri), tampak banyak gangguan spektral berupa efek bintik-bintik acak (*salt-and-pepper*) di mana perkebunan sawit dan hutan terfragmentasi secara artifisial oleh piksel lahan terbuka akibat variasi kabut lokal dan bayangan awan kumululus. Sebaliknya, pada peta hasil klasifikasi citra terkoreksi HistSpec (sebelah kanan), batas zonasi lanskap terlihat jauh lebih koheren, batas spasial kanopi hutan alami dan perkebunan sawit terpetakan secara tegas, serta sebaran permukiman linier di sepanjang jalan protokol terpetakan secara utuh.

3.3 Pembahasan

Di kawasan Ibu Kota Nusantara (IKN), Prasetyo dkk. [1] melaporkan *Overall Accuracy* (OA) sebesar 85,47% dan koefisien Kappa 0,80 dengan skema klasifikasi empat kelas. Namun, penelitian tersebut mengandalkan fusi data Sentinel-1 SAR untuk membantu mengatasi keterbatasan citra optik akibat tutupan awan. Sementara itu, Al Fathoni dkk [2] memperoleh *Recall* 88% dan *F1-score* 87,99% pada lima kelas menggunakan komposit tahunan GEE. Walaupun efektif dalam mengurangi tutupan awan, pendekatan komposit masih berpotensi menyisakan pengaruh *haze* yang meningkatkan *spectral confusion* antara hutan dan semak belukar. Penelitian ini memperoleh OA 79,80% dan Kappa 0,7549 pada enam kelas tutupan lahan, dengan kemampuan membedakan hutan alami dan perkebunan kelapa sawit menggunakan citra



Sentinel-2 tanpa dukungan data radar. Nilai Kappa tersebut termasuk kategori *substantial agreement* dan menunjukkan bahwa model mampu memberikan klasifikasi yang cukup andal pada lanskap tropis yang kompleks. Perbedaan jumlah kelas dan tingkat kemiripan karakteristik spektral antarkelas menjadi faktor penting yang perlu dipertimbangkan ketika membandingkan capaian akurasi dengan penelitian terdahulu [3], [17]. Komparasi hasil penelitian dapat dilihat pada tabel 5.

Tabel 5. Komparasi Hasil Penelitian dengan Studi-Studi Terdahulu Terkait Klasifikasi Tutupan Lahan Berbasis Citra Sentinel-2 dan Machine Learning (2021–2025)

Peneliti & Tahun	Lokasi & Data Citra	Metode & Model	Jumlah Kelas	Akurasi (OA & Kappa)
Prasetyo dkk. (2025) [1]	BWP I IKN, Kaltim; Sentinel-1 & Sentinel-2	Random Forest (integrasi optik & radar SAR)	4 Kelas	OA: 85,47% Kappa: 0,80
Al Fathoni dkk. (2025) [2]	Kawasan IKN, Kaltim; Sentinel-2 MSI (GEE)	Random Forest via GEE komposit tahunan	5 Kelas	Recall 88% F1score: 87,99
Waladi (2025) [16]	Provinsi Jambi; Sentinel-2 MSI	Random Forest berbasis indeks vegetasi	9 Kelas	OA: 85,91% F1 score : 85,48%
Gülci dkk. (2025) [9]	Kawasan Mediterania; Sentinel-2 MSI	Random Forest vs CART (Pohon Tunggal)	5 Kelas	OA: 98% (RF) OA: 95% (CART)
Nugraha dkk. (2025) [15]	DAS Cimanuk, Jabar; Sentinel-2 multitemporal	Random Forest, SVM, & Gradient Boosting	7 Kelas	OA: 86,03,90% Kappa: 0,83
Aziz dkk. (2024)[10]	Kawasan Hutan Pegunungan; Citra Satelit Optik	Random Forest vs ANN	4 Kelas	OA (ANN): 97,77% Kappa: 0,965
Penelitian Ini (2026)	Kab. Penajam Paser Utara Sentinel-2 10 m	Histogram Spesifikasi + 9 Fitur + Random Forest (uji 7 ML)	6 Kelas	OA: 79,80% Kappa: 0,7549

Dapat dilihat pada Tabel 5 menampilkan pembeda utama penelitian ini adalah penerapan *Histogram Specification* (*HistSpec*) sebagai tahap normalisasi radiometrik [7]. Waladi [16] memperoleh OA 85,91% dan F1-score 85,48% pada sembilan kelas di Provinsi Jambi menggunakan indeks vegetasi, tetapi penelitian tersebut dilakukan pada bentang perkebunan kelapa sawit yang relatif homogen dan belum menerapkan penyesuaian distribusi radiometrik berbasis citra referensi. Pada penelitian ini, *HistSpec* meningkatkan OA dari 78,33% menjadi 79,80% atau sebesar 1,47%, serta meningkatkan Kappa dari 0,7367 menjadi 0,7549. Penyesuaian *Cumulative Distribution Function* (CDF) antara citra target dan citra referensi memungkinkan distribusi nilai spektral menjadi lebih konsisten tanpa mengubah geometri piksel [7]. Hasil ini menunjukkan bahwa normalisasi radiometrik dapat membantu mengurangi pengaruh atmosfer, khususnya *haze*, sehingga informasi spektral yang digunakan model menjadi lebih stabil.

Analisis *feature importance* juga menunjukkan bahwa saluran *Near-Infrared* (NIR/B8) dan *Enhanced Vegetation Index* (EVI) merupakan fitur yang paling berpengaruh. Kedua fitur tersebut berperan penting dalam membedakan hutan alami dan perkebunan kelapa sawit. Model *Random Forest* menghasilkan *Producer's Accuracy* (PA) sebesar 60,98% dan *User's Accuracy* (UA) sebesar 69,44% untuk hutan alami, serta PA 76,74% dan UA 80,49% untuk perkebunan kelapa sawit. Temuan ini sejalan dengan penelitian Shandu dkk. [3] dan Dobrinić dkk[8] yang menunjukkan pentingnya informasi NIR dalam menangkap perbedaan struktur dan karakteristik vegetasi. EVI juga menunjukkan kontribusi lebih besar dibandingkan NDVI didukung temuan Vélez dkk [21] dan Radočaj dkk [20] bahwa EVI lebih mampu mempertahankan sensitivitas pada vegetasi yang rapat sekaligus mengurangi pengaruh aerosol atmosferik.

Dari perspektif spasial, peta tutupan lahan enam kelas yang dihasilkan memberikan *baseline geospasial* untuk memahami kondisi tutupan lahan di wilayah penyangga IKN, khususnya Kabupaten Penajam Paser Utara. Hasil klasifikasi menunjukkan dominasi hutan alami sebesar 35,60%, diikuti perkebunan kelapa sawit sebesar 26,90%, semak/vegetasi budidaya sebesar 14,80%, dan lahan terbuka sebesar 13,50%. Komposisi tersebut menunjukkan pentingnya pemantauan perubahan tutupan lahan, terutama pada wilayah yang mengalami tekanan pembangunan dan perubahan penggunaan lahan. Dengan resolusi spasial 10 meter dan hasil klasifikasi yang relatif bersih dari *noise salt-and-pepper*, peta yang dihasilkan dapat dimanfaatkan sebagai data awal untuk mendukung pemantauan perubahan lahan dan perencanaan tata ruang berkelanjutan di kawasan IKN.

4. KESIMPULAN

Penelitian ini telah berhasil mengkaji dan membuktikan efektivitas integrasi metode pra-pemrosesan *Histogram Specification* dengan algoritma pembelajaran mesin *Random Forest* untuk klasifikasi tutupan lahan 6 kelas berbasis citra Sentinel-2 di Kabupaten Penajam Paser Utara. Transformasi *Histogram Specification* terbukti mampu meratakan kontras



radiometrik, mengeliminasi distorsi kabut atmosferik tropis, serta menekan *noise salt-and-pepper* pada ruang fitur spektral. Dari evaluasi komparatif terhadap tujuh algoritma *Machine Learning*, *Random Forest* memberikan performa terbaik dan paling konsisten dengan *Overall Accuracy* sebesar 79,80% (naik dari 79,31% pada skenario RAW), Macro F1 sebesar 81,32%, koefisien Cohen's Kappa 0,7549 (*Substantial Agreement*), serta efisiensi waktu komputasi yang lincah (123 ms). Analisis feature importance membuktikan bahwa saluran inframerah dekat (B8 / NIR) dan indeks EVI memberikan kontribusi paling dominan dengan skor 0,198 dan 0,126, dalam membedakan spektral antara hutan alami dan perkebunan kelapa sawit. Selain itu, penelitian ini berhasil mengungkap kelemahan arsitektur *Gradient Boosting* yang rentan mengalami *overfitting* pada citra optik mentah (turun -4,92%), menegaskan keunggulan komparatif *Random Forest* untuk pemetaan penginderaan jauh regional. Peta tematik tutupan lahan 6 kelas yang diproduksi menampilkan basis data spasial yang akurat dan obyektif guna mendukung pengawasan tata ruang, pengendalian deforestasi, serta kelestarian ekologis di kawasan penyangga Ibu Kota Nusantara.

REFERENCES

- [1] D. Prasetyo, F. Luvian Fatah, N. Bashit, and F. Hadi, "Integrasi Citra Sentinel-1 Dan Sentinel-2 Untuk Pemetaan Tutupan Lahan Tahun 2024," *Jurnal Geodesi dan Geomatika*, vol. 08, no. 01, 2025, doi: 10.14710/elipsoida.2025.27043.
- [2] H. Al Fathoni, A. Junaidi, and F. P. Aditiawan, "Klasifikasi Tutupan Lahan Pada Citra Sentinel-2 Di Kawasan Ikn Menggunakan Google Earth Engine," 2025. doi: 10.36040/jati.v9i3.13652.
- [3] I. D. Shandu, S. Xulu, and M. Gebreslasie, "Enhancing land cover classification in the heterogeneous landscape by integrating auxiliary data with Sentinel-2 imagery using the random forest algorithm," *Frontiers in Remote Sensing*, vol. 6, 2026, doi: 10.3389/frsen.2025.1697897.
- [4] D. Phiri, M. Simwanda, S. Salekin, V. R. Nyirenda, Y. Murayama, and M. Ranagalage, "Sentinel-2 data for land cover/use mapping: A review," Jul. 01, 2020, *MDPI AG*. doi: 10.3390/rs12142291.
- [5] O. Danylo *et al.*, "A map of the extent and year of detection of oil palm plantations in Indonesia, Malaysia and Thailand," *Sci. Data*, vol. 8, no. 1, Dec. 2021, doi: 10.1038/s41597-021-00867-1.
- [6] W. Zhang, X. Yang, Z. Yuan, Z. Chen, and Y. Xu, "A Framework for Fine-Grained Land-Cover Classification Using 10 m Sentinel-2 Images," *Remote Sens. (Basel)*, vol. 16, no. 2, Jan. 2024, doi: 10.3390/rs16020390.
- [7] J. Cheng, J. Dai, and T. Zhang, "Radiometric resolution transformation of remote sensing based on histogram matching," in *Sixth International Conference on Geoscience and Remote Sensing Mapping (GRSM 2024)*, Z. Qin, J. Chen, and H. Wu, Eds., SPIE, 2025, p. 1350628. doi: 10.1117/12.3057823.
- [8] D. Dobrinić, M. Gašparović, and D. Medak, "Sentinel-1 and 2 time-series for vegetation mapping using random forest classification: A case study of northern Croatia," *Remote Sens. (Basel)*, vol. 13, no. 12, Jun. 2021, doi: 10.3390/rs13122321.
- [9] S. Gülcü, M. Wing, and A. E. Akay, "Land Use and Land Cover (LULC) Mapping Accuracy Using Single-Date Sentinel-2 MSI Imagery with Random Forest and Classification and Regression Tree Classifiers," *Geomatics*, vol. 5, no. 3, Sep. 2025, doi: 10.3390/geomatics5030029.
- [10] G. Aziz, N. Minallah, A. Saeed, J. Frnda, and W. Khan, "Remote sensing based forest cover classification using machine learning," *Sci. Rep.*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-023-50863-1.
- [11] G. De Luca, J. M. N. Silva, S. Di Fazio, and G. Modica, "Integrated use of Sentinel-1 and Sentinel-2 data and open-source machine learning algorithms for land cover mapping in a Mediterranean region," *Eur. J. Remote Sens.*, vol. 55, no. 1, pp. 52–70, 2022, doi: 10.1080/22797254.2021.2018667.
- [12] T. Zhang, J. Su, Z. Xu, Y. Luo, and J. Li, "Sentinel-2 satellite imagery for urban land cover classification by optimized random forest classifier," *Applied Sciences (Switzerland)*, vol. 11, no. 2, pp. 1–17, Jan. 2021, doi: 10.3390/app11020543.
- [13] G. Allen, R. De Veaux, and R. Nugent, *An Introduction to Statistical Learning with Applications in R*. Springer, 2021.
- [14] M. Vizzari, "PlanetScope, Sentinel-2, and Sentinel-1 Data Integration for Object-Based Land Cover Classification in Google Earth Engine," *Remote Sens. (Basel)*, vol. 14, no. 11, Jun. 2022, doi: 10.3390/rs14112628.
- [15] F. Nugraha, D. P. T. Baskoro, and S. D. Tarigan, "Evaluasi Algoritma Machine Learning untuk Klasifikasi dan Prediksi Penggunaan Lahan," *Majalah Geografi Indonesia*, vol. 39, no. 1, pp. 19–28, Jun. 2025, doi: 10.22146/mgi.70636.
- [16] A. Waladi, "Peningkatan Akurasi Klasifikasi Tutupan Lahan Menggunakan Random Forest pada Data Sentinel-2 di Jambi," *Jurnal Fasikom*, vol. 15, Apr. 2025, doi: 10.37859/jf.v15i1.8886.
- [17] R. G. Congalton and K. Green, *Assessing the Accuracy of Remotely Sensed Data Principles and Practices, Third Edition*, 3rd ed. CRS Press Taylor & Francis Group, 2019.
- [18] H. Hidayat, E. Azzahra, and F. Bioresita, "Analisa Perubahan Tutupan Lahan Akibat Letusan Gunung Semeru Tahun 2021 dengan Algoritma Random Forest (Studi Kasus: Kecamatan Pronojiwo, Kabupaten Lumajang)," *Geoid Journal of Geodesy and Geomatics*, vol. 19, no. 2, pp. 236–245, 2024, doi: 10.12962/geoid.v19i2.1179.
- [19] S. Cuypers, A. Nascetti, and M. Vergauwen, "Land Use and Land Cover Mapping with VHR and Multi-Temporal Sentinel-2 Imagery," *Remote Sens. (Basel)*, vol. 15, no. 10, May 2023, doi: 10.3390/rs15102501.
- [20] D. Radočaj, A. Šiljeg, R. Marinović, and M. Jurišić, "State of Major Vegetation Indices in Precision Agriculture Studies Indexed in Web of Science: A Review," Mar. 01, 2023, *MDPI*. doi: 10.3390/agriculture13030707.
- [21] S. Vélez, R. Martínez-Peña, and D. Castrillo, "Beyond Vegetation: A Review Unveiling Additional Insights into Agriculture and Forestry through the Application of Vegetation Indices," *J (Basel)*, vol. 6, no. 3, pp. 421–436, Jul. 2023, doi: 10.3390/j6030028.