

Klasifikasi Multi-Kelas Kanker Paru Berbasis Ekstraksi Fitur Hibrida GLCM, LBP dan Gabor Filter Menggunakan Algoritma *Random Forest* dan KNN pada Citra CT-Scan

Muhammad Iqbal, Panji Triadyaksa*, Qidir Maulana Binu Soesanto
Departemen Fisika, Fakultas Sains dan Matematika, Universitas Diponegoro

Info Artikel

Histori Artikel:

Diajukan: 12 Agustus 2025
Direvisi: 20 Desember 2025
Diterima: 30 Januari 2026

Kata kunci:

AUC
CT-Scan
Fitur Hibrida
Kanker Paru
KNN

Keywords:

AUC
CT-Scan
Hybrid Feature
Lung Cancer
KNN

Penulis Korespondensi:

Pandji Triadyaksa
Email:
p.triadyaksa@fisika.fsm.undip.ac.id

ABSTRAK

Penelitian ini bertujuan meningkatkan kinerja klasifikasi multi-kelas kanker paru jenis *Non-Small Cell Lung Cancer* (NSCLC) pada citra CT-Scan dengan menggunakan metode ekstraksi fitur hibrida serta algoritma klasifikasi *Random Forest* dan *K-Nearest Neighbor* (KNN). Dataset berasal dari Kaggle dengan total 851 citra, terdiri atas 309 citra *adenocarcinoma*, 156 citra *large cell carcinoma*, 242 citra *squamous cell carcinoma*, dan 144 citra normal. Dataset dibagi menjadi 70% data latih dan 30% data uji tanpa augmentasi tambahan. Pra-pemrosesan dilakukan dengan melakukan resize dan peningkatan kontras menggunakan *Contrast Limited Adaptive Histogram Equalization* (CLAHE). Ekstraksi fitur menggunakan *Gray Level Co-occurrence Matrix* (GLCM), *Local Binary Pattern* (LBP), dan Gabor Filter dengan panjang gelombang 3-40 piksel dan enam orientasi ($0^\circ - 150^\circ$). Kombinasi fitur tekstur kemudian diklasifikasikan menggunakan algoritma *Random Forest* dan KNN. Hasil menunjukkan kombinasi GLCM+LBP dengan *Random Forest* memberikan performa terbaik dengan akurasi 99,22%, dan AUC 1,00, sedangkan kombinasi Gabor Filter+GLCM dengan KNN menghasilkan akurasi 91,37%, dan AUC 0,94. Nilai AUC > 0,90 pada seluruh kelas menunjukkan kedua model memiliki kemampuan diskriminatif yang sangat baik. Penelitian ini menyimpulkan bahwa kombinasi fitur tekstur hibrida dengan *Random Forest* dan KNN mampu menghasilkan model klasifikasi yang akurat, efisien, serta berpotensi untuk dikembangkan sebagai sistem pendukung diagnosis otomatis dalam bidang fisika medis.

This study aims to improve the multi-class classification performance of Non-Small Cell Lung Cancer (NSCLC) on CT-scan images by employing a hybrid feature extraction approach with Random Forest and K-Nearest Neighbor (KNN) classification algorithms. The dataset was obtained from Kaggle, consisting of a total of 851 images, including 309 adenocarcinoma, 156 large cell carcinoma, 242 squamous cell carcinoma, and 144 normal lung images. The dataset was divided into 70% training and 30% testing data without additional augmentation. Preprocessing was performed by resizing and enhancing image contrast using Contrast Limited Adaptive Histogram Equalization (CLAHE). Feature extraction employed Gray Level Co-occurrence Matrix (GLCM), Local Binary Pattern (LBP), and Gabor Filter with wavelengths ranging from 3 to 40 pixels and six orientations ($0^\circ - 150^\circ$). The combined texture features were then classified using Random Forest and KNN algorithms. The results showed that the combination of GLCM+LBP with Random Forest achieved the best performance with an accuracy of 99.22% and an AUC of 1.00, while the combination of Gabor Filter+GLCM with KNN achieved 91.37% accuracy and 0.94 AUC. An AUC value greater than 0.90 for all classes indicates that both models have excellent discriminative capability. This study concludes that the hybrid texture feature combination with Random Forest and KNN produces an accurate and efficient classification model, with strong potential for development as an automatic diagnostic support system in medical physics.

Copyright © 2026 Author(s). All rights reserved

I. PENDAHULUAN

Salah satu penyebab utama kematian akibat kanker di dunia adalah kanker paru. Di Indonesia, kanker paru berada di peringkat ketiga setelah kanker payudara dan kanker serviks dengan angka kasus baru 8,8%, serta menjadi penyebab kematian tertinggi akibat kanker yaitu 13,2% dari jumlah total kasus (Kementerian Kesehatan, 2023). Angka mortalitas yang tinggi ini disebabkan oleh kendala utama yaitu gejala awal yang tidak spesifik, sehingga diagnosis sering terlambat. Diagnosis yang masih dilaksanakan dengan metode konvensional seperti biopsi memiliki sifat invasif dan membutuhkan waktu yang relatif lama, sehingga perlu adanya pendekatan lain seperti pencitraan medis yang lebih akurat dan efisien.

CT-Scan merupakan salah satu solusi penting dalam memberikan gambaran struktur paru-paru secara detail dan membantu membedakan jenis kanker paru (Wahyuni & Amalia, 2022). Kanker paru dibagi menjadi dua kategori yaitu *Small Cell Lung Cancer* (SCLC) dan *Non-Small Cell Lung Cancer* (NSCLC). NSCLC mencakup sekitar 80-85% dari seluruh kasus kanker paru dan terdiri atas tiga sub tipe utama, yaitu *adenocarcinoma*, *squamous cell carcinoma*, dan *large cell carcinoma* (Wasilah et al., 2023). Pada umumnya *adenocarcinoma* muncul di jaringan perifer paru dan dapat bermetastasis ke organ lain, *squamous cell carcinoma* cenderung berkembang di bagian sentral bronkus, sedangkan *large cell carcinoma* memiliki ciri sel abnormal yang berukuran besar dan prognosis yang buruk (Ozdemir et al., 2020; Siegel et al., 2019). Oleh karena itu, klasifikasi multi-kelas sub tipe NSCLC memiliki nilai klinis yang sangat penting dalam rangka mendukung diagnosis dan penentuan terapi.

Berbagai penelitian terdahulu telah menunjukkan efektivitas teknik pengolahan citra dan *machine learning* dalam klasifikasi kanker paru. Li et al. (2024) telah menunjukkan bahwa kombinasi *Gray Level Co-occurrence Matrix* (GLCM), Haralick, dan *Autoencoder* dengan *Support Vector Machine* (SVM) dapat mencapai akurasi hingga 100%. Penelitian lain yang dilakukan Nair et al. (2024) memperoleh akurasi 99,65% dengan menggabungkan GLCM, *Wavelet*, *shape*, dan intensitas menggunakan algoritma *Random Forest*. Kemudian Kumari et al. (2025) yang menggabungkan *Local Binary Pattern* (LBP), *Local Directional Pattern* (LDP), *Discrete Wavelet Transform* (DWT), dan Gabor Filter juga mencapai akurasi 100% menggunakan *Random Forest*. Hasil-hasil penelitian tersebut telah menjadi bukti bahwa beragam ekstraksi fitur yang digunakan mampu menghasilkan citra medis yang lebih komprehensif.

Pada penelitian lainnya, beberapa peneliti menekankan penggunaan algoritma *K-Nearest Neighbor* (KNN) dalam klasifikasi kanker paru. Sathishkumar et al. (2019) mengusulkan sistem deteksi kanker paru menggunakan kombinasi pra-pemrosesan, segmentasi dan ekstraksi fitur GLCM untuk mengidentifikasi nodul kanker. Pada tahap klasifikasi, penelitian tersebut menggunakan algoritma SVM dan KNN untuk membedakan citra normal dan kanker. Kemudian Bhuvaneswari dan Therese (2015) mengusulkan metode deteksi kanker paru menggunakan teknik ekstraksi fitur Gabor Filter dan klasifikasi KNN yang dioptimasi dengan *Genetic Algorithm* (GA), menghasilkan akurasi hingga 90%. Penelitian-penelitian ini menunjukkan bahwa metode berbasis ekstraksi fitur manual dan algoritma klasifikasi tradisional seperti KNN tetap memberikan performa baik untuk deteksi kanker paru. Namun demikian, fokus penelitian masih pada klasifikasi biner dan belum banyak yang mengklasifikasikan multi-kelas sub tipe NSCLC.

Permasalahan ini mulai diatasi oleh Pahlevi et al. (2025) yang melakukan klasifikasi sub tipe NSCLC berbasis citra CT-Scan menggunakan GLCM dengan algoritma *Extreme Learning Machine* (ELM), dengan dukungan pra-pemrosesan *Histogram Equalization*. Akurasi maksimum yang diperoleh dalam penelitian ini sebesar 91,06%. Meskipun demikian, penelitian ini masih memiliki keterbatasan karena hanya memanfaatkan satu jenis ekstraksi fitur, sehingga informasi tekstur terbatas dan berimbas pada performa model. Urgensi klasifikasi multi-kelas sub tipe NSCLC ditegaskan kembali oleh Zhang et al. (2025) mengekstrak lebih dari 1.800 fitur radiomik dari citra CT dan menguji pada 11 algoritma klasifikasi, penelitian ini berhasil mencapai nilai *Area Under Curve* (AUC) 0,99-1,00 pada *cross-validation*. Namun, performa model menurun signifikan saat diuji pada data independen dengan AUC 0,61-0,64.

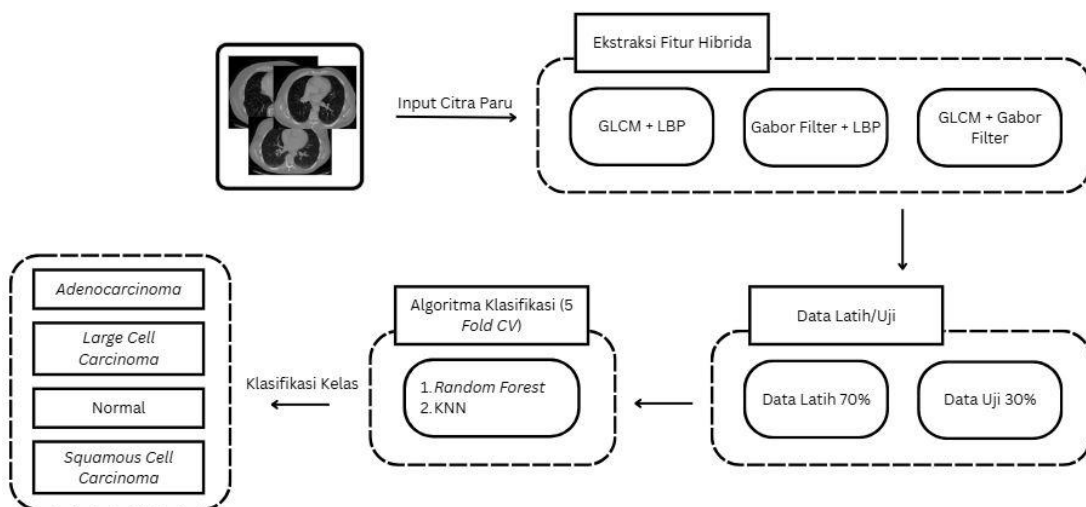
Untuk mengatasi keterbatasan tersebut, penelitian ini hadir dengan mengusulkan metode ekstraksi fitur hibrida yang mengkombinasikan GLCM, LBP, dan Gabor Filter. GLCM telah menjadi ekstraksi fitur tekstur yang sering digunakan untuk analisis tekstur berbasis distribusi spasial intensitas piksel (Salunkhe et al., 2021). LBP merupakan metode tekstur lokal sederhana berbasis perbedaan intensitas piksel tetangga (Song et al., 2013). Kemudian Gabor Filter sering digunakan untuk

menganalisis komponen frekuensi spasial dan orientasi (Rai & Rivas, 2020). Fitur tekstur dari citra CT kemudian diklasifikasikan menggunakan dua algoritma *machine learning*, yaitu *Random Forest* dan KNN. *Random Forest* dipertimbangkan karena kemampuannya dalam menangani fitur tekstur hibrida yang banyak dan kompleks, sementara KNN dipilih karena kesederhanaannya dalam menangani pola lokal antar kelas pada dataset berukuran kecil hingga menengah. Evaluasi kinerja model dilakukan dengan menggunakan metrik seperti akurasi, sensitivitas, spesifisitas, dan AUC yang telah umum digunakan dalam menilai kinerja model klasifikasi (Grandini et al., 2020; ID, 2021).

Dengan demikian, penelitian ini bertujuan untuk meningkatkan kinerja model klasifikasi multi-kelas subtype NSCLC berbasis citra CT-Scan melalui pendekatan ekstraksi fitur hibrida (GLCM, LBP, dan Gabor Filter) dengan algoritma *Random Forest* dan KNN. Kebaruan penelitian ini terletak pada penerapan kombinasi metode ekstraksi fitur untuk klasifikasi empat kelas citra CT-Scan paru, yang terdiri tiga subtype NSCLC (*Adenocarcinoma*, *Large Cell Carcinoma*, dan *Squamous Cell Carcinoma*) serta kelas normal. Berbeda dengan sebagian besar penelitian terdahulu yang masih terbatas pada klasifikasi biner atau hanya menggunakan satu jenis fitur tekstur, penelitian ini secara sistematis mengevaluasi efektivitas kombinasi pasangan fitur tekstur (GLCM+LBP, Gabor+LBP, dan GLCM+Gabor) dalam representasi citra medis. Selain itu, mengacu pada temuan Zhang et al. (2025) yang menunjukkan penurunan performa saat diuji pada data independen, penelitian ini berupaya mengatasi keterbatasan tersebut dengan kombinasi fitur yang lebih representatif untuk meningkatkan kemampuan generalisasi model klasifikasi multi-kelas NSCLC. Hasil penelitian ini harapannya dapat berkontribusi dalam pengembangan sistem klasifikasi kanker paru berbasis *machine learning* yang lebih akurat, efisien, dan dapat digeneralisasi pada data klinis yang lebih beragam. Dengan kemampuan mengklasifikasi subtype NSCLC, penelitian ini berpotensi membantu proses diagnosis yang lebih tepat sehingga dapat menurunkan angka mortalitas akibat kanker paru.

II. METODE

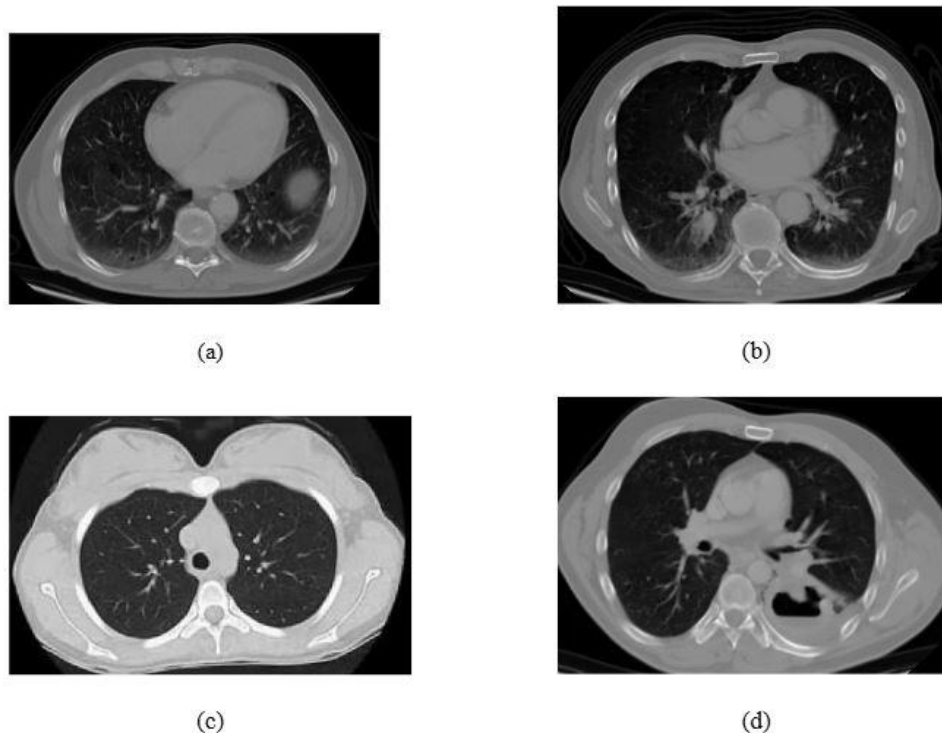
Penelitian ini dilaksanakan di Laboratorium Radiasi, Departemen Fisika, Fakultas Sains dan Matematika, Universitas Diponegoro. Metode penelitian ini disusun sedemikian rupa dengan beberapa tahapan, mulai dari pengumpulan dataset, pra-pemrosesan, ekstraksi fitur, penggabungan fitur, klasifikasi, validasi, dan evaluasi performa. Seluruh tahapan penelitian diimplementasikan menggunakan perangkat lunak MATLAB versi R2024a (Mathworks Inc., USA) dengan lisensi institusional resmi milik Universitas Diponegoro. Secara umum, skematik metode penelitian klasifikasi multi-kelas kanker paru ditunjukkan pada Gambar 1, di mana gambar ini menggambarkan proses dari tahap input citra hingga klasifikasi.



Gambar 1 Diagram skematik penelitian

2.1 Dataset

Penelitian ini menggunakan data yang berasal dari dataset publik dari platform Kaggle <https://www.kaggle.com/datasets/borhanitrash/lung-cancer-ct-scan-dataset>, berupa citra medis CT-Scan paru-paru. Total citra medis pada dataset yang digunakan adalah 851 dengan distribusi kelas *adenocarcinoma* (309 citra), *large cell carcinoma* (156 citra), normal (144 citra), dan *squamous cell carcinoma* (242 citra). Kemudian citra ini diberi label numerik kelas 1 untuk *adenocarcinoma*, kelas 2 untuk *large cell carcinoma*, kelas 3 untuk normal, dan kelas 4 untuk *squamous cell carcinoma*. Contoh citra untuk masing-masing kelas ditampilkan pada Gambar 2.



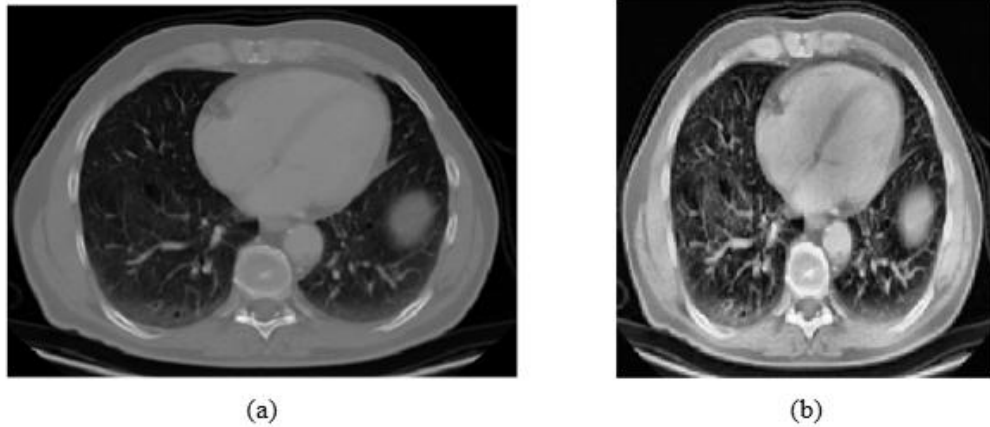
Gambar 2 Citra CT-Scan untuk masing-masing kelas (a) *Adenocarcinoma* (b) *Large Cell Carcinoma* (c) Normal (d) *Squamous Cell Carcinoma*

Seluruh citra digunakan dalam proses *training* dan *testing*, di mana data latih untuk *training* sebanyak 70% dan data uji untuk *testing* sebanyak 30%. Pada penelitian ini tidak dibuat subset data validasi secara terpisah, karena proses validasi dilakukan secara internal melalui teknik *k-fold Cross-Validation* pada data latih. Pendekatan ini umum digunakan untuk dataset berukuran kecil hingga menengah agar mencegah berkurangnya stabilitas model (Lei & Li, 2025; Mehan, 2025; Nahiduzzaman et al., 2024; Zhang et al., 2025). Dengan demikian, data uji 30% berfungsi sebagai *independent hold-out set* yang tidak terlibat dalam proses *training* maupun validasi, sehingga dapat digunakan untuk menilai kemampuan generalisasi model pada data yang baru.

2.2 Pra-Pemrosesan Citra

Tahap pra—pemrosesan dilakukan untuk meningkatkan kualitas citra agar nantinya mudah dilakukan analisis dan ekstraksi fitur-fitur. Citra medis yang digunakan perlu dilakukan standarisasi dalam aspek dimensi, di mana dalam penelitian ini seluruh citra diubah ukurannya menjadi 200 x 200 piksel menggunakan teknik *resize* di MATLAB. Ukuran ini dipilih karena efisiensi komputasi dan ketepatan representasi struktur paru (Sinha et al., 2025). Ukuran citra yang lebih kecil dari ini berpotensi menghilangkan detail tekstur halus, sedangkan ukuran yang lebih besar dapat meningkatkan waktu komputasi tanpa memberikan peningkatan model klasifikasi yang berarti. Proses *resize* dilakukan secara proporsional agar tidak terjadi distorsi bentuk anatomi paru. Kemudian agar citra memiliki kontras yang lebih baik, digunakan teknik *Contrast Limited Adaptive Histogram* (Clahe). Teknik Clahe bekerja dengan mengurangi *noise* melalui pemotongan histogram pada ambang batas yang telah ditentukan (*clip limit*), serta mendistribusikan ulang nilai intensitas piksel (Nahiduzzaman et al., 2024). Perbedaan

tampilan citra asli dengan citra yang sudah dilakukan *resize* dan Clahe ditunjukkan pada Gambar 3. Perbedaan tampilan ini hanyalah efek visual akibat penyesuaian skala intensitas dan peningkatan kontras, bukan perubahan bentuk atau struktur morfologis paru secara anatomi.



Gambar 3 Contoh hasil pra-pemrosesan citra CT-Scan paru-paru kelas *adenocarcinoma* (a) citra asli (b) citra hasil *Resize + Clahe*

2.3 Ekstraksi Fitur

2.3.1 *Gray Level Co-occurrence Matrix* (GLCM)

Metode *Gray Level Co-occurrence Matrix* (GLCM) digunakan untuk menganalisis hubungan antar piksel dalam citra berdasarkan intensitas keabuan. Citra diubah ke format 8-bit (0–255) menggunakan `im2uint8` dan dihitung matriks *Co-occurrence* pada empat orientasi sudut (0° , 45° , 90° , dan 135°) dengan jarak antar piksel 1 menggunakan fungsi `graycomatrix`. Pemilihan jarak 1 piksel didasarkan pada pertimbangan bahwa hubungan spasial antar piksel yang berdekatan lebih mampu merepresentasikan pola tekstur lokal yang sensitif terhadap variasi mikrostruktur jaringan paru. Penggunaan nilai jarak yang kecil juga disarankan karena dapat menangkap detail halus tanpa kehilangan informasi lokal (Gadkari, 2004). Dari setiap matriks, dihitung empat parameter tekstur utama, yaitu *Contrast*, *Correlation*, *Energy*, dan *Homogeneity* menggunakan fungsi `graycoprops`. Nilai dari keempat arah dirata-ratakan untuk membentuk vektor fitur berdimensi 1×4 .

$$f_1 = \sum_{n=0}^{N_g-1} n^2 \left(\sum_j \sum_j M_{ij} \right) \quad (1)$$

$$f_2 = \sum_i \sum_j \frac{ij M_{ij} - \mu_x \mu_y}{\sigma_x \sigma_y} \quad (2)$$

$$f_3 = \sum_i \sum_j M_{ij}^2 \quad (3)$$

$$f_4 = \sum_i \sum_j \frac{M_{ij}}{1 + |i - j|} \quad (4)$$

Persamaan statistik dari fitur GLCM ditunjukkan pada persamaan (1) sampai (4). Nilai f_1 adalah *contrast*, f_2 adalah *correlation*, f_3 adalah *energy*, dan f_4 adalah *homogeneity*. Pada persamaan tersebut ada beberapa simbol, di mana N_g adalah jumlah tingkat keabuan, i adalah nomor baris, j adalah nomor kolom, M_{ij} adalah elemen ke- i, j dari GLCM yang telah dinormalisasi. Kemudian μ adalah rata-rata GLCM dan σ adalah variasi dari intensitas seluruh piksel referensi dalam hubungan yang berkontribusi terhadap GLCM (Salunkhe et al., 2021).

2.3.2 Local Binary Pattern (LBP)

Metode *Local Binary Pattern* (LBP) berfungsi untuk mendeskripsikan pola tekstur lokal dengan membandingkan intensitas piksel pusat terhadap tetangganya. Implementasi dilakukan menggunakan fungsi `extractLBPFeatures` dengan parameter radius = 1, jumlah tetangga = 8, dan ukuran sel = [16 16]. Konfigurasi LBP dengan radius bernilai satu dan jumlah tetangga delapan dipilih karena mewakili versi *circularly symmetric* yang mampu mendeskripsikan pola tekstur lokal secara rotasi-invarian (Ojala et al., 2002). Sedangkan parameter ukuran sel [16 16] dipilih berdasarkan pertimbangan keseimbangan kekayaan pola tekstur dan kestabilan histogram fitur. Ukuran sel yang terlalu kecil (misal [8 8]) menghasilkan pola LBP yang terlalu sedikit sehingga kurang representatif, jika ukuran sel terlalu besar (misal [32 32]) menyebabkan *smoothing* berlebihan yang dapat mengaburkan variasi tekstur lokal (Wang et al., 2009). Nilai fitur LBP dihasilkan dalam bentuk histogram yang menggambarkan distribusi pola biner lokal pada setiap citra.

2.3.3 Filter Gabor

Filter Gabor digunakan untuk mengekstraksi fitur tekstur berbasis frekuensi spasial. Filter ini diterapkan menggunakan fungsi `imgaborfilt` dengan panjang gelombang yang diatur secara logaritmik antara 3 hingga 40 piksel dan enam orientasi sudut (0°, 30°, 60°, 90°, 120°, dan 150°). Pemilihan enam orientasi tersebut dimaksudkan untuk merepresentasikan variasi pola tekstur dalam seluruh arah dominan dengan interval 30°, sehingga mampu menangkap perbedaan orientasi struktur jaringan paru. Hasil respon filter Gabor menggambarkan struktur tekstur citra pada berbagai arah dan skala.

2.4 Penggabungan Fitur Tekstur

Penggabungan fitur tekstur yang diperoleh dari kombinasi beberapa metode ekstraksi dilakukan untuk meningkatkan performa klasifikasi. Tiga model hibrida diuji, yaitu (1) GLCM + LBP, (2) Gabor Filter + LBP, (3) Gabor Filter + GLCM. Vektor fitur dari masing-masing model akan digabungkan secara horizontal untuk membentuk representasi citra yang kaya terhadap variasi pola tekstur.

2.5 Klasifikasi

2.5.1 K-Nearest Neighbors (KNN)

Algoritma KNN bekerja berdasarkan jarak terdekat antara data uji dan data latih. Parameter yang digunakan adalah *Number of Neighbors* adalah satu, *distance metric* adalah *Euclidean*, dan *distance weight* adalah *equal*. Seluruh fitur dinormalisasi untuk memastikan skala antar fitur seragam.

2.5.2 Random Forest

Algoritma *Random Forest* menggunakan pendekatan *Bagged Trees* dengan parameter utama *maximum number of splits* adalah 595, *number of learners* adalah 30. Model ini dipilih karena kemampuannya menangani data berdimensi tinggi dan mengurangi risiko *overfitting*. Pada penelitian ini, hasil ekstraksi fitur hibrida (GLCM, LBP, dan Gabor Filter) menghasilkan sekitar 8.500 fitur numerik untuk masing-masing citra. Jumlah fitur yang jauh lebih besar dibandingkan jumlah sampel (851 citra) menyebabkan data berdimensi tinggi, yang berpotensi menimbulkan penurunan kemampuan model untuk mengenali pola karena ruang data menjadi sangat luas. Dalam konteks tersebut, *Random Forest* memiliki mekanisme berupa *feature bagging* yang mampu melakukan klasifikasi secara stabil meskipun jumlah dimensi fitur yang tinggi.

2.6 Validasi dan Evaluasi Kinerja Model

Dataset dibagi menjadi 70% data latih dan 30% data uji. Proses validasi model dilakukan secara internal pada data latih dengan teknik *5-fold Cross-Validation* di dalam *Classification Learner* Matlab. Validasi silang ini digunakan dengan tujuan memvalidasi model agar terhindar *overfitting* dan memperoleh performa yang stabil sebelum dilakukan pengujian akhir pada data uji. Evaluasi model dilakukan menggunakan *Confusion Matrix*, dengan metrik utama yaitu akurasi, sensitivitas, spesifisitas, dan *area under curve* (AUC).

Persamaan (5), (6) dan (7) merupakan persamaan akurasi, sensitivitas, dan spesifisitas secara berurutan. TP adalah *True Positive*, TN adalah *True Negative*, FP adalah *False Positive*, dan FN adalah *False Negative*. Semakin tinggi nilai akurasi, spesifisitas, dan sensitivitas, maka semakin baik performa

model klasifikasi. Nilai AUC juga menjadi indikator penting yang menunjukkan kemampuan model dalam membedakan kelas secara menyeluruh.

$$A = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$Se = \frac{TP}{TP + FN} \quad (6)$$

$$Sp = \frac{TN}{TN + FP} \quad (7)$$

III. HASIL DAN DISKUSI

3.1 Perbandingan Kinerja Model Klasifikasi

Hasil evaluasi kinerja model pada Tabel 1 menunjukkan bahwa kombinasi fitur hibrida GLCM+LBP dengan algoritma *Random Forest* memberikan performa paling tinggi dengan akurasi 99,22%, sensitivitas 99,30%, spesifisitas 99,72%, dan nilai AUC 1,00. Nilai tersebut menunjukkan kemampuan model dalam mengklasifikasi empat kelas citra CT-Scan paru secara hampir sempurna. Hasil penelitian ini melampaui capaian penelitian Pahlevi et al. (2025) yang hanya mencapai akurasi 91,06% pada GLCM tunggal. Hal ini menunjukkan bahwa penggabungan fitur tekstur global GLCM dan mikrotekstur lokal LBP mampu menangkap variasi spasial serta struktural jaringan paru, dibanding hanya mengandalkan satu fitur tekstur tunggal.

Tabel 1 Hasil evaluasi model klasifikasi berdasarkan fitur hibrida dan algoritma KNN serta *Random Forest*

Kombinasi Fitur Hibrida	Algoritma Klasifikasi	Akurasi (%)	Sensitivitas (%)	Spesifisitas (%)	AUC
GLCM + LBP	KNN	90,20	89,81	96,47	0,93
GLCM + LBP	<i>Random Forest</i>	99,22	99,30	99,72	1,00
Gabor Filter + LBP	KNN	88,63	90,13	96,10	0,93
Gabor Filter + LBP	<i>Random Forest</i>	74,90	73,52	90,64	0,93
Gabor Filter + GLCM	KNN	91,37	90,46	96,89	0,94
Gabor Filter + GLCM	<i>Random Forest</i>	82,35	81,12	93,47	0,96

Keunggulan hasil ini sejalan dengan temuan Li et al. (2024) yang melaporkan akurasi mendekati 100% pada kombinasi GLCM, Haralick, dan *Autoencoder* dengan algoritma berbasis optimasi parameter. Namun, pendekatan tersebut memerlukan konfigurasi komputasi tinggi dan hanya mengklasifikasi dua kelas. Dalam hal ini, penelitian ini berhasil mengatasi keterbatasan dengan pendekatan fitur hibrida (GLCM+LBP) yang jauh lebih ringan secara komputasi dan sudah dapat membedakan empat kelas dari citra CT paru.

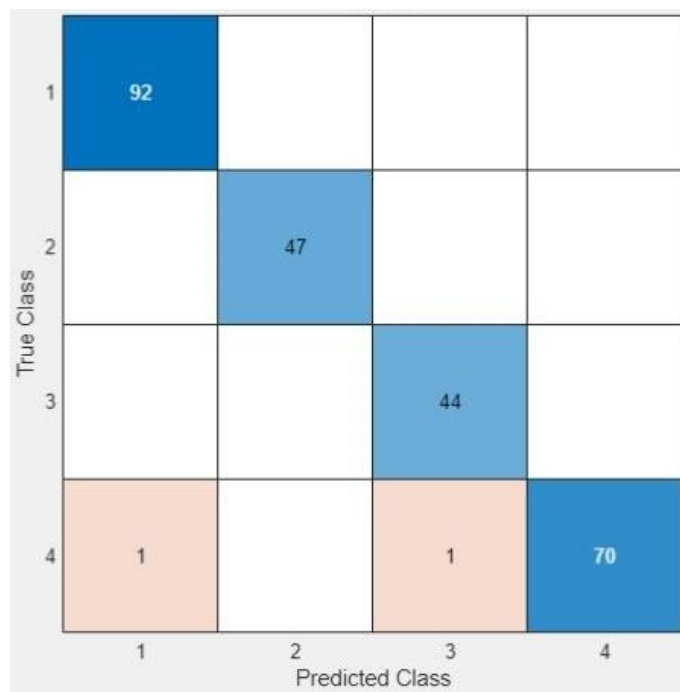
Kombinasi fitur yang dikombinasikan dengan Gabor Filter pada Tabel 1, menunjukkan performa yang lebih rendah dibandingkan dengan GLCM + LBP. Akurasi maksimum yang diperoleh dari kombinasi Gabor Filter adalah 91,37% pada algoritma klasifikasi KNN, sedangkan pada *Random Forest* akurasi lebih rendah yaitu 82,35%. Penurunan akurasi ini menunjukkan bahwa fitur frekuensi spasial dari Gabor Filter dapat menimbulkan redundansi informasi jika tidak diimbangi oleh mekanisme seleksi fitur yang proporsional. Fenomena ini juga diamati oleh Kumari et al. (2025) yang menekankan bahwa kombinasi fitur multi-level seperti LBP, Gabor Filter, DWT memang memperkaya informasi tekstur. Akan tetapi, informasi tekstur yang kaya akan dapat menurunkan stabilitas model tanpa optimasi selektif. Menurut Rai & Rivas (2020) setiap filter pada Gabor hanya mampu menangkap satu pola spesifik, sehingga untuk tekstur kompleks seperti jaringan paru yang memiliki variasi kontras dan orientasi beragam, diperlukan multiset filter dengan konfigurasi orientasi dan panjang gelombang yang tepat. Sehingga apabila parameter tidak selaras dengan arah dominan tekstur citra CT, sebagian informasi spasial penting dapat terlewat yang akan mengakibatkan penurunan akurasi klasifikasi.

Selain itu, kompleksitas frekuensi tinggi yang dihasilkan Gabor dapat menimbulkan korelasi antar fitur yang kurang sesuai dengan algoritma *ensemble Random Forest*, karena algoritma ini sangat sensitif terhadap fitur yang saling tumpang tindih. Berbanding terbalik dengan algoritma KNN yang

mengklasifikasi berdasarkan jarak antar vektor fitur tanpa melakukan proses pembelajaran eksplisit, sehingga lebih toleran terhadap fitur yang memiliki korelasi tinggi atau distribusi non-linear. Perbedaan akurasi yang dicapai KNN dan *Random Forest* saat menggunakan metode Gabor Filter pada Tabel 1 menunjukkan adanya perbedaan dalam pengolahan fitur tekstur. KNN cenderung lebih stabil ketika menangani variasi pola tekstur kompleks yang dihasilkan oleh frekuensi spasial seperti Gabor. Hal ini terjadi karena klasifikasi KNN dilakukan berdasarkan kedekatan pola fitur, bukan berdasarkan pembentukan batas keputusan global seperti pada *Random Forest*. Dengan demikian, dapat dikatakan kinerja KNN yang lebih baik dalam kombinasi Gabor Filter bukan karena superioritas algoritmanya, tetapi karena karakteristik Gabor Filter yang berorientasi spasial lebih sesuai dengan pendekatan berbasis jarak yang dilakukan oleh KNN.

3.2 Analisis Performa *Random Forest* dengan Kombinasi Ekstraksi Fitur Hibrida Terbaik

Visualisasi kinerja model GLCM+LBP+*Random Forest* pada Gambar 4 menunjukkan *confusion matrix* dengan tingkat ketepatan klasifikasi yang sangat tinggi pada seluruh kelas. Model ini mampu mengklasifikasi 92 citra *adenocarcinoma*, 47 citra *large cell carcinoma*, 44 citra normal, dan 70 citra *squamous cell carcinoma* secara benar. Kesalahan klasifikasi hanya dua yang terjadi pada kelas *squamous cell carcinoma* yang masing-masing teridentifikasi sebagai *adenocarcinoma* dan *large cell carcinoma*.



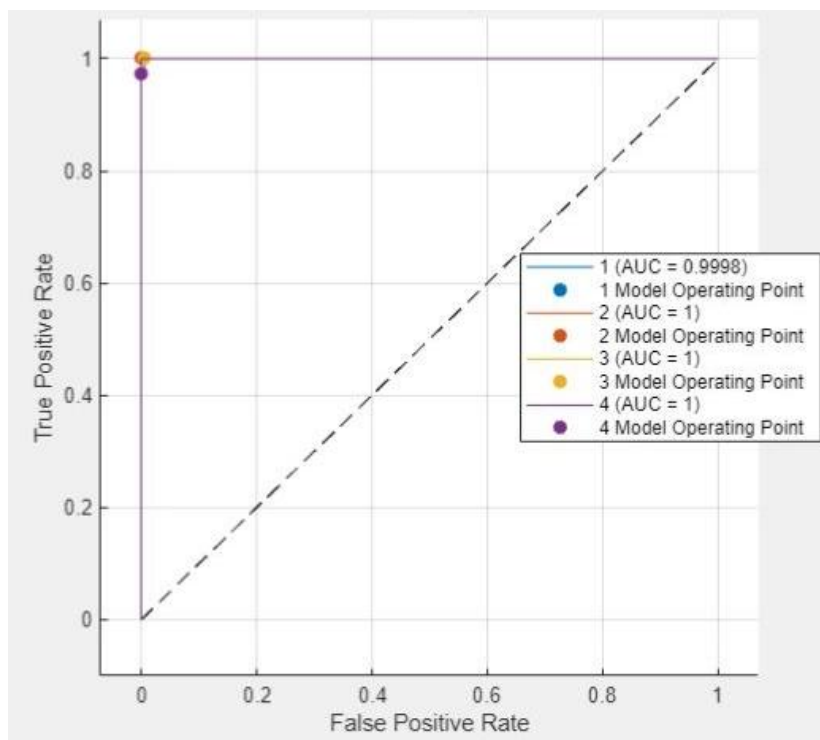
Gambar 4 *Confusion Matrix* GLCM + LBP algoritma *Random Forest*

Kinerja model ini menghasilkan tingkat akurasi, sensitivitas, dan spesifisitas yang baik, sehingga menunjukkan keseimbangan antara kemampuan mendeteksi jaringan kanker dan mempertahankan ketepatan terhadap jaringan normal. Nilai diagonal yang mendominasi matriks menandakan konsistensi internal model dan minimnya bias antar kelas. Keberhasilan model mengenali *adenocarcinoma* (92/92) dan *large cell carcinoma* (47/47) menunjukkan bahwa model mampu mengekstraksi pola tekstur halus dan distribusi intensitas khas dari sel dan jaringan. Pada umumnya, kedua jenis kanker ini memiliki perbedaan tekstur yang jelas terhadap jaringan paru normal, sehingga mudah dikenali oleh fitur berbasis statistik dan lokal. Kesalahan kecil yang muncul pada kelas *squamous cell carcinoma* kemungkinan disebabkan oleh kemiripan pola densitas dan arah serta jaringan dengan *adenocarcinoma* dan *large cell carcinoma*. Namun, meskipun terjadi sedikit *misclassification*, model tetap mempertahankan sensitivitas tinggi terhadap kelas tersebut 99,3%.

Pada Gambar 5, kurva *Receiver Operating Characteristic* (ROC) menunjukkan hasil evaluasi AUC untuk empat kelas. Nilai AUC yang berada sangat dekat atau tepat pada 1,00 menunjukkan bahwa model memiliki kemampuan diskriminatif yang hampir sempurna dalam membedakan tiap kelas kanker

paru dan jaringan normal. Hal ini memperkuat hasil *confusion matrix* pada Gambar 4, di mana kesalahan klasifikasi hanya terjadi pada dua citrai *squamous cell carcinoma*. Capaian ini memperkuat potensi model sebagai *Computer-Aided Diagnosis* (CAD) yang mampu membantu identifikasi dini berbagai subtype kanker paru berbasis CT-Scan dengan tingkat keyakinan klinis yang tinggi.

Analisis mendalam terhadap model GLCM+LBP menunjukkan bahwa performa tinggi *Random Forest* tidak hanya disebabkan oleh kekuatan klasifikasinya, tetapi juga oleh stabilitas hasil pada seluruh *fold* validasi. Dengan menggunakan *5-fold cross-validation*, model menunjukkan konsistensi akurasi di atas 99% menandakan kemampuan generalisasi yang baik dan minim *overfitting*. Pendekatan ini sejalan dengan metode validasi pada penelitian Mehan (2025), yang juga menekankan pentingnya *cross-validation* untuk dataset medis berukuran terbatas. Jika dibandingkan dengan penelitian Mehan (2025), yang menggunakan *transfer learning SqueezeNet* sebagai ekstraktor fitur, model yang diusulkan dalam penelitian ini memberikan peningkatan. Mehan melaporkan akurasi tertinggi 94,8% dan AUC 0,989 pada klasifikasi tiga kelas (*benign*, *malignant*, dan *normal*), sedangkan dalam model penelitian ini menghasilkan 99,22% akurasi untuk empat kelas. Perbedaan ini menegaskan bahwa fitur statistik berbasis tekstur seperti GLCM dan LBP mampu menghasilkan representasi yang lebih stabil terhadap variasi jaringan paru dibandingkan dengan fitur hasil *transfer learning* CNN seperti *SqueezeNet*.



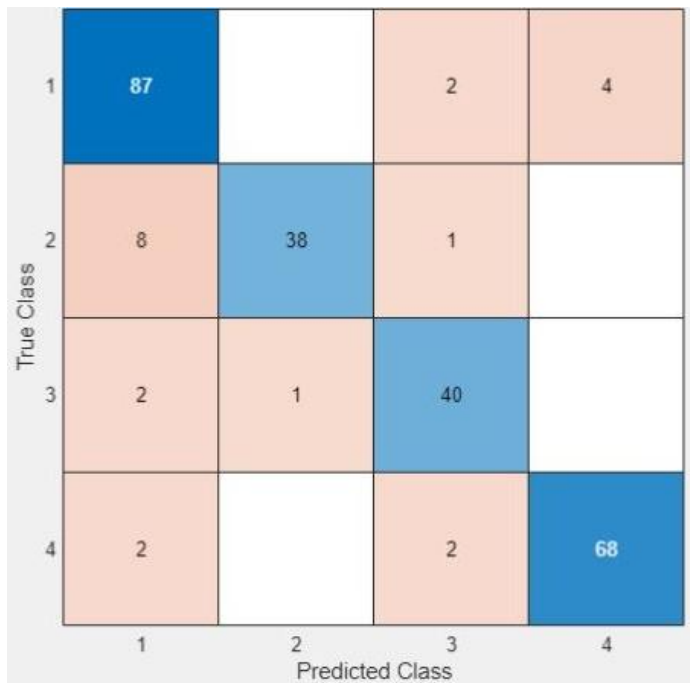
Gambar 5 AUC algoritma *Random Forest* untuk membedakan kelas 1 “*adenocarcinoma*”, kelas 2 “*large cell carcinoma*”, kelas 3 “*normal*”, kelas 4 “*squamous cell carcinoma*”

Hasil penelitian ini juga dapat dibandingkan dengan studi Nair et al. (2024) yang mengembangkan metode RWI dengan kombinasi ANN dan *Random Forest*. Dalam penelitian tersebut, dilaporkan bahwa model berbasis RWI-*Random Forest* mencapai akurasi 99,6%. Temuan tersebut mendukung hasil penelitian ini, di mana algoritma *Random Forest* juga menunjukkan kinerja optimal dengan akurasi 99,22%. Namun, Nair et al. (2024) hanya membedakan dua kelas utama, sedangkan penelitian ini berhasil melakukan klasifikasi empat kelas. Dengan demikian, meskipun tingkat akurasi penelitian Sneha lebih tinggi secara numerik, konteks multi-kelas pada penelitian ini menunjukkan kompleksitas dan daya generalisasi yang lebih besar.

3.3 Analisis Performa KNN dengan Kombinasi Ekstraksi Fitur Hibrida Terbaik

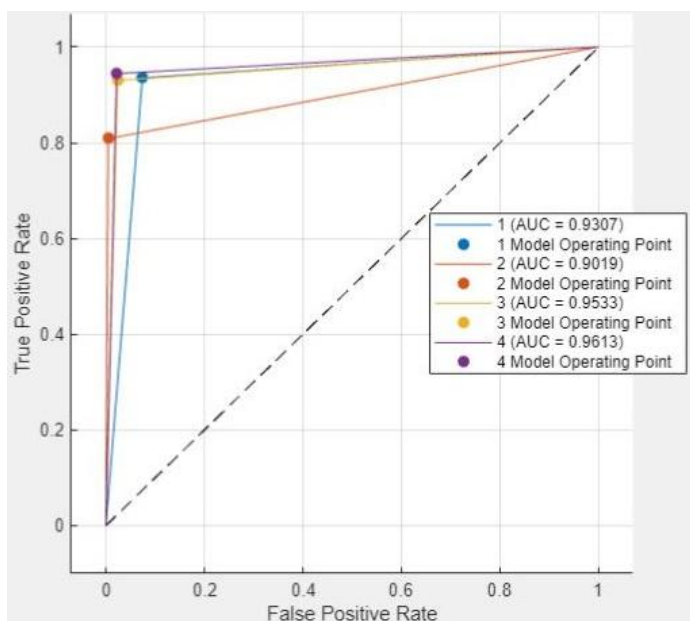
Berdasarkan Tabel 1 kombinasi ekstraksi fitur hibrida terbaik untuk model klasifikasi KNN adalah Gabor Filter + GLCM. Visualisasi kinerja model Gabor Filter + GLCM + KNN dapat dilihat menggunakan *confusion matrix* yang ditunjukkan pada Gambar 6. Model berhasil mengenali 87 citra

adenocarcinoma, 38 citra *large cell carcinoma*, 40 citra normal, dan 68 citra *squamous cell carcinoma* secara benar. Kesalahan klasifikasi sebagian besar terjadi antar kelas yang memiliki kemiripan pola tekstur, seperti antara *adenocarcinoma* dengan *large cell carcinoma*, serta antara normal dengan *squamous cell carcinoma*. Kinerja model ini berhasil mencapai akurasi 91,37%, sensitivitas 90,46%, spesifisitas 96,89%, dan AUC 0,94 ketika diaplikasikan pada data uji/testing.



Gambar 6 Confusion Matrix Gabor Filter + GLCM algoritma KNN

Kurva ROC untuk model klasifikasi KNN dengan ekstraksi fitur Gabor Filter dan GLCM dapat dilihat pada Gambar 7. Berdasarkan grafik ROC, seluruh kelas menunjukkan nilai AUC di atas 0,90, yang menandakan bahwa model mampu melakukan klasifikasi dengan sangat baik. Kelas *squamous cell carcinoma* memiliki nilai AUC tertinggi yaitu 0,96 yang artinya model paling efektif dalam mengenali pola tekstur kelas tersebut. Sedangkan nilai AUC terendah dimiliki oleh kelas *large cell carcinoma* yaitu 0,90, hal ini menandakan bahwa model masih melakukan kesalahan dalam mengenali pola tekstur yang sedikit tumpang tindih dengan kelas lain.



Gambar 7 AUC algoritma KNN untuk membedakan kelas 1 “*adenocarcinoma*”, kelas 2 “*large cell carcinoma*”, kelas 3 “*normal*”, kelas 4 “*squamous cell carcinoma*”

Secara prinsip, KNN bekerja dengan menentukan kelas data uji berdasarkan jarak terdekat dengan data latih di mana dalam penelitian ini menggunakan *Euclidean* sebagai metrik jarak, dengan nilai $k = 1$. Pemilihan parameter ini relevan dengan data hasil ekstraksi Gabor Filter dan GLCM yang memiliki pola kompleks namun terpisah jelas antar kelas, sehingga *nearest neighbor* tunggal sudah cukup representatif. Jika dibandingkan dengan penelitian Bhuvaneswari dan Therese (2015) yang mengusulkan metode *Genetic KNN* berbasis Gabor Filter untuk deteksi kanker paru, hasil penelitian ini menunjukkan adanya peningkatan akurasi. Bhuvaneswari dan Therese juga menggunakan ekstraksi fitur Gabor Filter yang kemudian dioptimasi dengan GA untuk memilih tetangga terdekat secara adaptif. Penelitian tersebut melaporkan akurasi maksimum 90% pada klasifikasi dua kelas (normal vs kanker).

Penelitian ini menggunakan fitur hibrida yaitu Gabor Filter dan GLCM, tetapi tanpa proses optimasi genetika dan berhasil mencapai akurasi 91,37%. Kemudian penelitian ini melakukan klasifikasi empat kelas yaitu tiga subtype NSCLC serta normal. Perbedaan penelitian ini dengan Bhuvaneswari dan Therese (2015) menunjukkan bahwa fitur hibrida mampu memberikan representasi tekstur yang lebih kaya dibandingkan hanya menggunakan Gabor Filter tunggal. Selain itu, meskipun penelitian ini tidak melibatkan proses optimasi seperti GA, hasil akurasi yang dicapai justru lebih tinggi.

IV. KESIMPULAN

Penelitian ini menunjukkan bahwa penerapan kombinasi fitur tekstur hibrida GLCM dan LBP yang diklasifikasikan menggunakan algoritma *Random Forest* mampu memberikan performa paling optimal dalam identifikasi multi-kelas kanker paru berbasis citra CT-Scan. Model yang dihasilkan memiliki kemampuan yang sangat baik dalam membedakan subtype *adenocarcinoma*, *large cell carcinoma*, *squamous cell carcinoma*, dan jaringan paru normal dengan akurasi 99,22%, sensitivitas 99,30%, spesifisitas 99,72% dan nilai AUC 1,00. Pendekatan fitur tekstur hibrida Gabor Filter dan GLCM yang diklasifikasikan menggunakan algoritma KNN juga berhasil menunjukkan akurasi yang baik yaitu 91,37%, sensitivitas, spesifisitas, dan nilai AUC. Hal ini menunjukkan bahwa model KNN tetap memiliki potensi baik dalam menangani pola tekstur kompleks. Jika dilakukan perbandingan kedua model antara *Random Forest* dan KNN, terlihat bahwa algoritma *ensemble Random Forest* lebih unggul dalam stabilitas dan generalisasi hasil, sedangkan KNN lebih adaptif terhadap variasi spasial dari fitur tekstur Gabor Filter.

Secara keseluruhan, hasil penelitian ini membuktikan bahwa penggabungan fitur tekstur (GLCM, LBP, dan Gabor Filter) dengan pendekatan hibrida mampu menghasilkan representasi citra yang lebih kaya daripada hanya mengandalkan satu jenis fitur tekstur. Pendekatan fitur tekstur hibrida yang dikombinasikan dengan algoritma *Random Forest* dan KNN juga terbukti dalam meningkatkan akurasi klasifikasi multi-kelas subtype NSCLC serta berpotensi dalam sistem CAD di bidang fisika medis. Hanya saja penelitian ini belum menggunakan tahap seleksi fitur yang akan menyaring fitur tekstur paling relevan. Sehingga saran penelitian selanjutnya adalah menambahkan tahap seleksi fitur sebagai langkah untuk mengurangi dimensi fitur yang berlebih, menekan redundansi informasi, serta meningkatkan efisiensi komputer.

DAFTAR PUSTAKA

- Bhuvaneswari, P., & Therese, A. B. (2015). Detection of Cancer in Lung with K-NN Classification Using Genetic Algorithm. *Procedia Materials Science*, 10(Cnt 2014), 433–440. <https://doi.org/10.1016/j.mspro.2015.06.077>
- Gadkari, D. (2004). *Image Quality Analysis Using GLCM Image Quality Analysis Using GLCM STARS Citation STARS Citation*. <http://library.ucf.edu>
- Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for Multi-Class Classification: an Overview*. 1–17. <http://arxiv.org/abs/2008.05756>
- ID, I. D. (2021). *Machine Learning Teori, Studi Kasus dan Implementasi Menggunakan Python* (1 ed.). Badan Penerbit Universitas Riau.
- Kementerian Kesehatan, R. (2023). Panduan Penatalaksanaan Kanker Paru. *Komite Penanggulangan Kanker Nasional*, 10–11.

- Kumari, S., Jha, R., Ray, A., Jena, J. J., Gourisaria, M. K., & Bandyopadhyay, A. (2025). Explainable AI based Lung and Colon Cancer diagnosis using Enhanced Feature Analysis. *Procedia Computer Science*, 258, 18–29. <https://doi.org/10.1016/j.procs.2025.04.160>
- Lei, L., & Li, W. (2025). Transformer-based multi-task model for lung tumor segmentation and classification in CT images. *Journal of Radiation Research and Applied Sciences*, 18(3), 101657. <https://doi.org/10.1016/j.jrras.2025.101657>
- Li, L., Yang, J., Por, L. Y., Khan, M. S., Hamdaoui, R., Hussain, L., Iqbal, Z., Rotaru, I. M., Dobrotă, D., Aldrery, M., & Omar, A. (2024). Enhancing lung cancer detection through hybrid features and machine learning hyperparameters optimization techniques. *Heliyon*, 10(4). <https://doi.org/10.1016/j.heliyon.2024.e26192>
- Mehan, V. (2025). Advanced artificial intelligence driven framework for lung cancer diagnosis leveraging SqueezeNet with machine learning algorithms using transfer learning. *Medicine in Novel Technology and Devices*, 27(June). <https://doi.org/10.1016/j.medntd.2025.100383>
- Nahiduzzaman, M., Faisal Abdulrazak, L., Arselene Ayari, M., Khandakar, A., & Islam, S. M. R. (2024). A novel framework for lung cancer classification using lightweight convolutional neural networks and ridge extreme learning machine model with SHapley Additive exPlanations (SHAP). *Expert Systems with Applications*, 248(January). <https://doi.org/10.1016/j.eswa.2024.123392>
- Nair, S. S., Devi, V. N. M., & Bhasi, S. (2024). Enhanced lung cancer detection: Integrating improved random walker segmentation with artificial neural network and random forest classifier. *Heliyon*, 10(7). <https://doi.org/10.1016/j.heliyon.2024.e29032>
- Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7), 971–987. <https://doi.org/10.1109/TPAMI.2002.1017623>
- Ozdemir, O., Russell, R. L., & Berlin, A. A. (2020). A 3D Probabilistic Deep Learning System for Detection and Diagnosis of Lung Cancer Using Low-Dose CT Scans. *IEEE Transactions on Medical Imaging*, 39(5), 1419–1429. <https://doi.org/10.1109/TMI.2019.2947595>
- Pahlevi, O., Handrianto, Y., Ayu, D., Wulandari, N., Leidiyana, H., & Kanti, L. (2025). Klasifikasi Kanker Paru pada Citra CT Scan menggunakan Extreme Learning Machine dan Histogram Equalization. *Jurnal Informatika dan Rekayasa Perangkat Lunak*, 7(1).
- Rai, M., & Rivas, P. (2020). A Review of Convolutional Neural Networks and Gabor Filters in Object Recognition. *Proceedings - 2020 International Conference on Computational Science and Computational Intelligence, CSCI 2020*, 1560–1567. <https://doi.org/10.1109/CSCI51800.2020.00289>
- Salunkhe, S., Bachute, M., Gite, S., Vyas, N., Khanna, S., Modi, K., Katpatal, C., & Kotecha, K. (2021). Classification of alzheimer's disease patients using texture analysis and machine learning. *Applied System Innovation*, 4(3). <https://doi.org/10.3390/asi4030049>
- Sathishkumar, R., Kalaiarasan, K., Prabhakaran, A., & Aravind, M. (2019). Detection of Lung Cancer using SVM Classifier and KNN Algorithm. *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, 1–7. <https://doi.org/10.1109/ICSCAN.2019.8878774>
- Siegel, R. L., Miller, K. D., & Jemal, A. (2019). Cancer statistics, 2019. *CA: A Cancer Journal for Clinicians*, 69(1), 7–34. <https://doi.org/10.3322/caac.21551>
- Sinha, S., Mali, S., & Rajkhowa, S. (2025). High-Precision Lung Cancer Classification with Custom CNN: Evaluation and Transfer Learning for Broader Cancer Types. *Procedia Computer Science*, 258, 3922–3933. <https://doi.org/10.1016/j.procs.2025.04.644>
- Song, K. C., Yan, Y. H., Chen, W. H., & Zhang, X. (2013). Research and perspective on local binary pattern. *Zidonghua Xuebao/Acta Automatica Sinica*, 39(6), 730–744. [https://doi.org/10.1016/s1874-1029\(13\)60051-8](https://doi.org/10.1016/s1874-1029(13)60051-8)
- Wahyuni, S., & Amalia, L. (2022). Perkembangan Dan Prinsip Kerja Computed Tomography (CT Scan). *GALENICAL: Jurnal Kedokteran dan Kesehatan Mahasiswa Malikussaleh*, 1(2), 88. <https://doi.org/10.29103/jkkmm.v1i2.8097>

- Wang, X., Han, T. X., & Yan, S. (2009). An HOG-LBP Human Detector with Partial Occlusion Handling. *Proceedings of the IEEE International Conference on Computer Vision, Iccv*, 32–39. <https://doi.org/10.1109/ICCV.2009.5459207>
- Wasilah, H., Limbong, M., Zahra, F., Syahrir, A., Mutia, L., Rayasari, F., Herawati, A. T., Annisa, F., Cusmarih, Rukmi, D. K., Purnamasari, F., Saprudin, Y. P. N., Muhlis, R., & Anderson, E. (2023). *Patofisiologi untuk Mahasiswa Keperawatan*. Yayasan Kita Menulis.
- Zhang, J., Wei, J., & Sun, R. (2025). Integrative machine learning analysis of radiomics for the non-invasive prediction of non-small cell lung cancer subtypes. *Journal of Radiation Research and Applied Sciences*, 18(3), 101757. <https://doi.org/10.1016/j.jrras.2025.101757>