

Perbandingan Kinerja Identifikasi Model VGG-19 Dengan Inception V3 Dalam Klasifikasi Penyakit Appendicitis

Dwi Prapita Sari*, Ilka Zufria

Fakultas Sains dan Teknologi, Ilmu Komputer, Universitas Islam Negeri Sumatra Utara, Medan, Indonesia

Email: ¹*dwiprapitasari02@email.com, ²ilkazufria@uinsu.ac.id

Email Penulis Korespondensi: dwiprapitasari02@email.com

Submitted 26-01-2025; Accepted 24-02-2026; Published 25-02-2026

Abstrak

Appendicitis merupakan kondisi kegawatdaruratan bedah yang memerlukan diagnosis cepat dan akurat. Namun, keterbatasan interpretasi citra ultrasonografi (USG) sering kali menimbulkan risiko kesalahan diagnosis, khususnya pada kondisi data medis yang terbatas. Penelitian ini bertujuan untuk menentukan model klasifikasi yang paling efektif sebagai sistem pendukung keputusan klinis dengan membandingkan dua arsitektur *transfer learning* berbasis *Convolutional Neural Network* (CNN), yaitu VGG-19 dan InceptionV3. Menggunakan dataset sebanyak 2.168 citra yang dibagi dengan rasio 70% data latih, 10% validasi, dan 20% data uji, model dievaluasi berdasarkan metrik akurasi, *precision*, *recall*, *F1-score*, dan *Area Under Curve* (AUC). Hasil pengujian menunjukkan bahwa InceptionV3 memberikan performa yang jauh lebih unggul dengan akurasi sebesar 0.9033%, *F1-score* kelas appendicitis 0.8946%, dan nilai AUC 0.9502%. Sebaliknya, VGG-19 hanya mencapai akurasi 0.8255% dengan tingkat *recall* kelas appendicitis yang sangat rendah, yaitu 0.8019%. Rendahnya nilai *recall* pada VGG-19 mengindikasikan adanya risiko tinggi terjadinya *missed diagnosis*. Kontribusi penelitian ini adalah memberikan rekomendasi model yang lebih andal dan efektif untuk mendukung proses identifikasi appendicitis berbasis kecerdasan buatan, terutama dalam menangani keterbatasan data klinis.

Kata Kunci: Appendicitis; Convolutional Neural Network; InceptionV3; VGG-19; Klasifikasi Citra; Ultrasonography

Abstract

Appendicitis is a surgical emergency that requires rapid and accurate diagnosis. However, limitations in ultrasound (USG) image interpretation often pose a risk of misdiagnosis, particularly in scenarios with limited medical data. This study aims to determine the most effective classification model for a clinical decision support system by comparing two transfer learning-based Convolutional Neural Network (CNN) architectures: VGG-19 and InceptionV3. Utilizing a dataset of 2,168 images split into 70% training, 10% validation, and 20% testing data, the models were evaluated using metrics such as accuracy, precision, recall, F1-score, and Area Under Curve (AUC). The results demonstrate that InceptionV3 delivered significantly superior performance, achieving an accuracy of 0.9033%, an F1-score of 0.8946% for the appendicitis class, and an AUC of 0.9502%. In contrast, VGG-19 only reached an accuracy of 0.8255%, with a recall for the appendicitis class as low as 0.8019%. The poor recall performance of VGG-19 indicates a high risk of missed diagnosis. This research contributes by recommending a more reliable and effective model to support AI-based appendicitis identification, specifically in limited data scenarios.

Keywords: Appendicitis; Convolutional Neural Network; InceptionV3; VGG-19; Image Classification; Ultrasonography

1. PENDAHULUAN

Penyakit appendicitis merupakan salah satu kondisi medis yang umum terjadi dan sering kali memerlukan penanganan segera. Secara global, penyakit ini paling sering menyerang individu usia 10–20 tahun dengan insidensi sekitar 90–100 kasus per 100.000 orang per tahun [1]. Appendicitis dikategorikan sebagai kegawatdaruratan bedah karena risiko komplikasi serius yang dapat terjadi akibat pecahnya usus buntu (*ruptur*), sehingga memerlukan diagnosis yang cepat dan akurat. Proses diagnosis tradisional masih sangat bergantung pada kemampuan klinisi, sehingga risiko salah diagnosis tetap tinggi [2]. Kesulitan ini semakin meningkat karena gejala appendicitis sering mirip dengan gangguan perut lain, seperti nyeri perut kanan bawah, mual, dan demam, sehingga membuat deteksi dini menjadi tantangan medis yang signifikan.

Seiring dengan perkembangan teknologi medis, pendekatan berbasis pengolahan citra telah muncul sebagai solusi untuk meningkatkan akurasi dan kecepatan diagnosis. Penggunaan citra medis seperti USG dan CT Scan memungkinkan visualisasi kondisi usus buntu secara non-invasif. Namun, interpretasi manual tetap rentan terhadap subjektivitas dan keterbatasan pengalaman tenaga medis. Oleh karena itu, pemanfaatan metode *deep learning* untuk analisis citra medis menjadi sangat relevan, karena mampu mengenali pola visual kompleks yang sulit diidentifikasi secara manual [3].

Dalam konteks ini, penelitian berbasis Convolutional Neural Network (CNN) menawarkan pendekatan yang lebih objektif dan efisien. CNN mampu mengekstraksi fitur spasial dari citra medis dan melakukan klasifikasi secara otomatis. Model CNN populer seperti VGG-19 dan InceptionV3 memiliki kemampuan dalam mengenali pola visual yang kompleks, meskipun dengan karakteristik yang berbeda. VGG-19 dikenal memiliki arsitektur yang sederhana namun jumlah parameternya besar, sedangkan InceptionV3 lebih efisien dalam penggunaan parameter dan mampu mengekstraksi fitur multi-skala dengan cepat. Pendekatan perbandingan kinerja berbagai arsitektur CNN berdasarkan akurasi dan efisiensi komputasi juga telah dilakukan oleh Zufria et al. (2025), yang menunjukkan bahwa setiap arsitektur memiliki keunggulan berbeda tergantung pada kompleksitas data dan kebutuhan implementasi [4].

Penelitian ini memilih untuk membandingkan kedua arsitektur tersebut dalam klasifikasi appendicitis, dengan tujuan menemukan model terbaik berdasarkan beberapa faktor penting, termasuk jumlah parameter, ukuran model, efisiensi komputasi, dan waktu *inference* (kecepatan prediksi) [5]. Parameter ini menjadi sangat penting dalam konteks

implementasi klinis, karena memungkinkan diagnosis cepat dan dapat diterapkan di fasilitas medis dengan sumber daya terbatas. Dengan membandingkan akurasi dan efisiensi kedua model, penelitian ini dapat memberikan panduan pemilihan arsitektur CNN yang optimal untuk deteksi appendicitis [6].

Penelitian sebelumnya oleh Baştuğ et al. (2024) menunjukkan bahwa model U-Net mampu melakukan segmentasi apendiks dengan akurasi tinggi, dengan nilai DSC rata-rata di atas 0,8. Keunggulan utama dari metode tersebut adalah otomatisasi proses segmentasi, yang memungkinkan integrasi ke dalam sistem pendukung keputusan klinis dan mempercepat diagnosis, khususnya di fasilitas yang kekurangan tenaga ahli radiologi [7]. Namun, penelitian tersebut fokus pada segmentasi organ, bukan pada klasifikasi kondisi sehat atau sakit, sehingga masih terdapat celah dalam penerapan untuk diagnosis otomatis appendicitis secara langsung [8].

Selain penelitian berbasis segmentasi, beberapa studi lain juga telah menerapkan CNN untuk klasifikasi penyakit berbasis citra medis dengan hasil akurasi yang tinggi, namun sebagian besar berfokus pada organ seperti paru-paru atau payudara, sementara penelitian khusus klasifikasi appendicitis berbasis USG masih relatif terbatas. Hal ini menunjukkan adanya research gap pada aspek perbandingan arsitektur CNN yang secara spesifik menargetkan klasifikasi appendicitis dengan mempertimbangkan akurasi sekaligus efisiensi komputasi dalam satu kerangka evaluasi terpadu [9].

Berdasarkan celah tersebut, penelitian ini memberikan kontribusi berbeda dengan menekankan perbandingan kinerja dua arsitektur CNN dalam klasifikasi penyakit appendicitis [10]. Berbeda dengan penelitian sebelumnya yang hanya menitikberatkan pada segmentasi atau satu model tunggal, studi ini secara sistematis membandingkan VGG-19 dan InceptionV3 menggunakan parameter evaluasi komprehensif meliputi akurasi, precision, recall, F1-score, AUC, jumlah parameter, ukuran model, serta waktu inference, sehingga memberikan dasar pengambilan keputusan yang lebih aplikatif dalam konteks klinis. Penelitian ini mengembangkan kriteria evaluasi yang lebih tepat, menargetkan prediksi dengan nilai minimal 0,8 untuk probabilitas kelas, dengan dua kategori utama: appendicitis dan non-appendicitis. Dengan pendekatan ini, penelitian tidak hanya menilai akurasi model, tetapi juga kemampuan model dalam mendeteksi kasus positif, efisiensi komputasi, dan ukuran model.

Relevansi penelitian ini semakin meningkat karena klasifikasi appendicitis berbasis citra medis masih relatif terbatas dibandingkan penyakit lain. Dengan fokus pada perbandingan kinerja VGG-19 dan InceptionV3, studi ini diharapkan dapat memberikan gambaran yang lebih jelas mengenai kelebihan dan keterbatasan masing-masing arsitektur, serta menentukan model terbaik untuk digunakan sebagai sistem pendukung keputusan klinis. Hasil penelitian ini juga memiliki potensi untuk meningkatkan efektivitas diagnosis appendicitis dan mempercepat penanganan pasien.

Berdasarkan uraian di atas, penelitian ini diangkat dengan judul “Perbandingan Kinerja Identifikasi Model VGG-19 Dengan InceptionV3 Dalam Klasifikasi Penyakit Appendicitis”, yang bertujuan untuk menilai, membandingkan, dan menemukan model CNN terbaik yang dapat diterapkan secara praktis dalam konteks medis. Studi ini diharapkan memberikan kontribusi nyata bagi dunia medis, khususnya dalam pengembangan sistem pendukung keputusan klinis berbasis citra medis. Kontribusi utama penelitian ini adalah menyediakan analisis komparatif yang komprehensif dan berbasis bukti mengenai performa dua arsitektur CNN populer pada klasifikasi appendicitis berbasis citra USG, sehingga dapat menjadi rujukan ilmiah dan teknis dalam pengembangan sistem pendukung keputusan klinis berbasis kecerdasan buatan.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini dilakukan dengan mengikuti beberapa tahapan yang sistematis untuk memastikan tujuan penelitian dapat dicapai. Tahapan penelitian dimulai dari perencanaan, pengumpulan data, pengolahan data, hingga evaluasi model. Tahap perencanaan meliputi identifikasi masalah, perumusan tujuan penelitian, dan pemilihan metode yang sesuai, yaitu penggunaan Convolutional Neural Network (CNN) dengan arsitektur VGG-19 dan InceptionV3. Perencanaan ini penting untuk menetapkan arah penelitian serta memastikan bahwa setiap tahapan dapat dilakukan secara konsisten dan dapat dipertanggungjawabkan [10]. Tahap berikutnya adalah pengumpulan data. Data yang digunakan merupakan dataset citra medis appendicitis dari platform Kaggle, terdiri dari citra USG dan CT Scan area abdomen dengan kondisi appendicitis positif dan negatif. Dataset ini berjumlah 500 gambar, yang selanjutnya dibagi menjadi data latih dan data uji. Pengumpulan data juga mencakup studi kepustakaan untuk meninjau penelitian terdahulu dan teori terkait pengolahan citra serta klasifikasi penyakit menggunakan CNN [11].

Setelah data tersedia, dilakukan tahap analisis dan pengolahan data (*data processing*). Proses ini mencakup *pre-processing* citra seperti resize (224×224 piksel untuk VGG-19, 299×299 piksel untuk InceptionV3), normalisasi nilai piksel, dan augmentasi citra (rotasi, flip, zoom) untuk meningkatkan variasi data. Selanjutnya, data digunakan untuk melatih model CNN dengan parameter seperti *learning rate*, *batch size*, dan jumlah *epoch* yang telah ditentukan. Setelah proses pelatihan selesai, dilakukan evaluasi model menggunakan data uji dengan metrik akurasi, presisi, *recall*, F1-score, serta *confusion matrix*. Tahap akhir adalah perbandingan hasil model VGG-19 dan InceptionV3. Selain membandingkan akurasi, penelitian ini juga menganalisis kecepatan prediksi, kemampuan mendeteksi kasus positif, dan efisiensi ukuran model. Hasil perbandingan ini akan memberikan gambaran model mana yang paling efektif sebagai pendukung keputusan klinis dalam diagnosis appendicitis. Gambar 1 di bawah menunjukkan ilustrasi kerangka penelitian secara keseluruhan.

Untuk memperjelas alur penelitian, tahapan penelitian secara sistematis dapat dirangkum sebagai berikut: (1) Identifikasi masalah dan studi literatur → (2) Pengumpulan dataset citra medis → (3) Pre-processing dan augmentasi data → (4) Pembagian data (train, validation, test) → (5) Implementasi transfer learning VGG-19 dan InceptionV3 → (6) Pelatihan model → (7) Evaluasi performa (accuracy, precision, recall, F1-score, AUC, confusion matrix) → (8) Analisis perbandingan dan penarikan kesimpulan.

2.2 Metode Penyelesaian Masalah

Metode penyelesaian masalah dalam penelitian ini dilakukan dengan pendekatan eksperimen berbasis CNN. Tahapan metode ini dimulai dengan identifikasi masalah berupa kesulitan diagnosis appendicitis secara cepat dan akurat, kemudian dilanjutkan dengan pengumpulan dataset citra medis. Dataset yang telah tersedia selanjutnya diproses menggunakan *pre-processing*, augmentasi, dan normalisasi untuk meningkatkan kualitas data sebelum pelatihan.

Tahap selanjutnya adalah implementasi model CNN. Model VGG-19 dan InceptionV3 dilatih dengan parameter yang sama agar perbandingan performa lebih adil. Evaluasi dilakukan dengan data uji menggunakan metrik seperti akurasi, presisi, *recall*, F1-score, serta analisis *confusion matrix*. Selain itu, pengukuran kecepatan prediksi dan ukuran model dilakukan untuk menilai efisiensi masing-masing arsitektur.

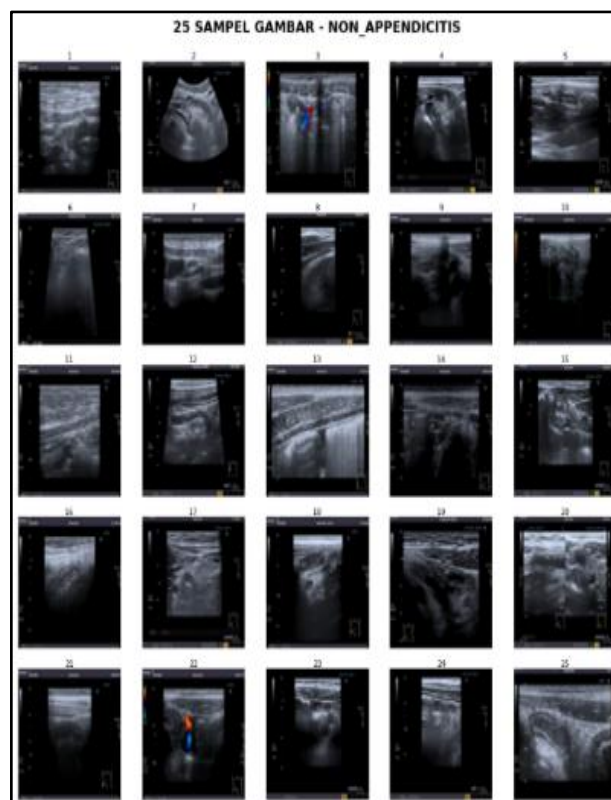
Hasil evaluasi model kemudian dibandingkan untuk menentukan model yang paling optimal. Kriteria perbandingan meliputi kemampuan mendeteksi kasus positif appendicitis, keseimbangan antara sensitivitas dan spesifisitas, serta efisiensi komputasi. Analisis ini diharapkan dapat memberikan panduan pemilihan model CNN yang sesuai untuk mendukung keputusan klinis dalam diagnosis penyakit appendicitis.

3. HASIL DAN PEMBAHASAN

Pada bagian ini, penulis membahas empat aspek utama, yaitu pengumpulan data, analisis data, representasi data, dan tahap awal perancangan model. Setiap sub-bab dijelaskan secara mendalam untuk memberikan pemahaman komprehensif mengenai penelitian yang dilakukan.

3.1 Pengumpulan Data

Penelitian ini menggunakan dataset citra medis kasus appendicitis dan non-appendicitis yang diperoleh dari platform Kaggle. Dataset terdiri dari citra ultrasonografi (USG) yang telah diklasifikasikan menjadi dua kelas: appendicitis dan non-appendicitis. Dataset berjumlah total 2.168 citra, masing-masing kelas berjumlah 1.084 citra agar seimbang. Data dibagi menjadi tiga subset: train, validasi, dan test, dengan proporsi sebagai berikut: subset train berjumlah 759 citra per kelas, subset validasi 113 citra per kelas, dan subset test 212 citra per kelas. Pembagian dilakukan secara acak sambil menjaga keseimbangan antar kelas.



Gambar 1. Sample Citra Appendicitis



Gambar 2. Sampel Citra Non-Appendicitis

3.2 Analisis Data

Analisis data bertujuan untuk membangun dua model klasifikasi citra menggunakan algoritma VGG-19 dan InceptionV3. Tahapan analisis dimulai dari persiapan dataset, praproses citra, hingga pembentukan arsitektur model. Dataset yang diperoleh diatur ulang ke dalam subset train, validation, dan test agar seimbang, sehingga model tidak bias terhadap salah satu kelas. Distribusi citra pada tiap subset dapat dilihat pada Tabel 1.

Tabel 1. Distribusi Jumlah Citra USG

Subset	Kelas	Jumlah Citra
Train	Appendicitis	759
Train	Non-Appendicitis	759
Valid	Appendicitis	113
Valid	Non-Appendicitis	113
Test	Appendicitis	212
Test	Non-Appendicitis	212
Total	Seluruh Citra	2.168

Berdasarkan tabel 1, setiap kelas, yaitu appendicitis dan non-appendicitis, memiliki jumlah citra yang seimbang pada setiap subset, yaitu 759 citra untuk pelatihan, 113 citra untuk validasi, dan 212 citra untuk pengujian. Keseimbangan jumlah data antar kelas ini sangat penting untuk mencegah bias model terhadap salah satu kelas, sehingga model dapat belajar secara adil dan menghasilkan performa klasifikasi yang lebih akurat dan objektif. Secara keseluruhan, total dataset yang digunakan dalam penelitian ini berjumlah 2.168 citra, yang memberikan representasi data yang cukup untuk melatih dan mengevaluasi kedua model CNN yang digunakan.

3.3 Representasi dan Praproses Data

Semua citra diolah menggunakan kelas ImageDataGenerator dari Keras untuk penyesuaian ukuran, normalisasi, dan augmentasi.

a. Resize Citra

1. Untuk VGG-19: 224×224 piksel
2. Untuk InceptionV3: 299×299 piksel
3. Citra diubah menjadi format tiga kanal (RGB).

b. Normalisasi

1. VGG-19: menggunakan fungsi `vgg19.preprocess_input` dengan mean subtraction ($R=123.68$, $G=116.779$, $B=103.939$)

2. InceptionV3: menggunakan fungsi `inception_v3.preprocess_input` agar piksel berada pada skala yang sesuai dengan bobot pra-latih ImageNet.
- c. Augmentasi Data (*hanya pada subset train*)
 1. Rotasi maksimum 15°
 2. Pergeseran lebar dan tinggi hingga 10%
 3. Zoom hingga 10%
 4. Horizontal flip
 5. `fill_mode='nearest'`



Gambar 3. Contoh Ilustrasi Augmentasi Rotasi

Gambar 3. menunjukkan contoh ilustrasi augmentasi rotasi yang akan lanjut ke tahap Grayscale

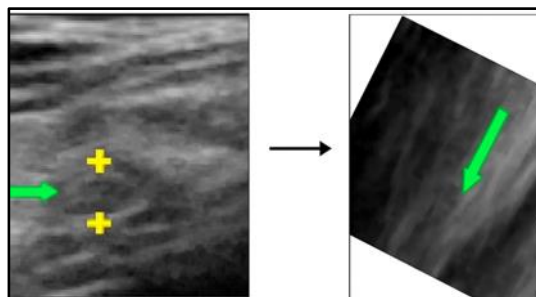
4. Konversi Grayscale ke RGB
Setiap nilai piksel grayscale direplikasi ke tiga channel (R, G, B). Contoh sampel 5×5 piksel setelah resize dan konversi dapat dilihat pada Gambar 4.7–4.9.
5. Rumus Resize (Bilinear Interpolation)

$$F_{(x',y')}^{(resize)} = (1 - \alpha)(1 - \beta)F_{(x,y)}^{(0)} + \alpha(1 - \beta)F_{(x+1,y)}^{(0)} + (1 - \alpha)\beta F_{(x,y+1)}^{(0)} + \alpha\beta F_{(x+1,y+1)}^{(0)}$$

Dimana $x = x' \times \frac{640}{224}$, $y = y' \times \frac{640}{224}$, $\alpha = \text{frac}(x)$, $\beta = \text{frac}(y)$. Contoh perhitungan untuk beberapa posisi piksel disajikan secara lengkap untuk memastikan akurasi transformasi citra.

3.4 Ringkasan Representasi VGG-19

Citra USG berukuran asli 640×640 piksel diresize menjadi 224×224 piksel, dikonversi ke RGB, dan dinormalisasi. Proses ini memastikan bahwa input citra siap untuk diproses oleh arsitektur VGG-19 sebagai feature extractor. Kualitas representasi citra yang baik berpengaruh langsung terhadap performa model dalam membedakan citra appendicitis dan non-appendicitis.



Gambar 4. Citra Asli 640×640 Grayscale

Berdasarkan, Gambar 4 menunjukkan representasi awal citra sebelum dilakukan transformasi lebih lanjut dalam tahapan preprocessing.

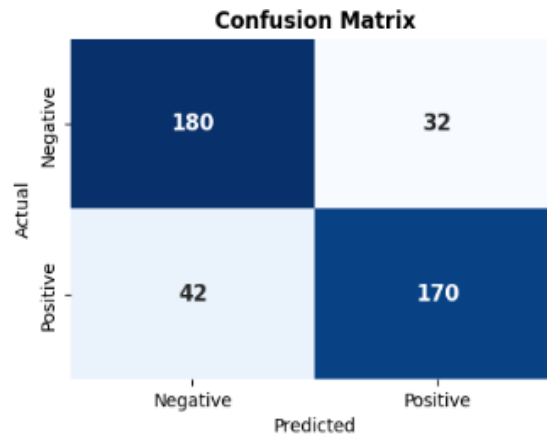
3.5 Perancangan

Berikut ini akan dibahas mengenai hasil yang telah didapatkan pada penelitian ini berupa pengujian yaitu sebagai berikut.

3.5.1 Hasil pengujian model VGG-19

Berdasarkan hasil pelatihan model VGG-19, terlihat bahwa perperforma pada data pelatihan terlihat stabil dengan nilai akurasi berada di kisaran 59-54%, serta nilai AUC yang relatif konsisten, yang menandakan model mampu mempelajari

pola pada data latih. Namun, perperforma tersebut tidak diikuti oleh peningkatan yang sesuai pada data validasi, dimana nilai validasi accuracy cenderung stagnan dan rendah di sekitar 50-54%, disertai dengan validasi loss yang tinggi dan tidak menunjukkan tren penurunan yang jelas. yaitu kondisi ketika model terlalu menyesuaikan diri terhadap data latih tetapi gagal melakukan generalisasi dengan baik pada data yang tidak terlihat sebelumnya. Oleh karena itu, dapat disimpulkan bahwa model VGG-19 kurang mampu digunakan pada dataset dalam penelitian ini karena belum mampu mencapai hasil yang baik.



Gambar 5. Confusion matrix vgg-19

Hasil *confusion matrix* pada model VGG-19 menunjukkan bahwa performa klasifikasi masih belum optimal. Dari total sampel kelas appendicitis, sebanyak 170 kasus berhasil diprediksi dengan benar, sementara 42 kasus lainnya salah diklasifikasikan sebagai non-appendicitis. Pada kelas non-appendicitis, model mampu mengidentifikasi 180 kasus dengan benar, namun masih terdapat 32 kasus yang keliru diprediksi sebagai appendicitis. Pola kesalahan ini menunjukkan bahwa VGG-19 belum mampu membedakan kedua kelas secara konsisten, sehingga sensitivitas dan spesifisitas model masih relatif rendah.

Berikut ini perhitungan matrix pada data validasi dalam klasifikasi penyakit Appendicitis sesuai kelasnya:

a. Accuracy (appendicitis)

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{170 + 180}{424} = \frac{350}{424} = 0.8255 = 82.55\%$$

Accuracy (Non-Appendicitis)

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{170 + 180}{424} = \frac{350}{424} = 0.8255 = 82.55\%$$

b. Precision(appendicitis)

$$precision = \frac{TP}{TP + FP} = \frac{170}{170 + 32} = \frac{170}{202} = 0.8416 = 84.16\%$$

precision (Non-appendicitis)

$$\frac{TP}{TP + FP} = \frac{180}{180 + 42} = \frac{180}{222} = 0.8108 = 81.08\%$$

c. Sensitivity/recall (appendicitis)

$$\frac{TP}{TP + FN} = \frac{170}{170 + 42} = \frac{170}{212} = 0.8019 = 80.19\%$$

Sensitivity/recall (Non-appendicitis)

$$\frac{TP}{TP + FN} = \frac{180}{180 + 32} = \frac{180}{212} = 0.8491 = 84.91\%$$

d. Specificity (appendicitis)

$$\frac{TN}{TN + FP} = \frac{180}{180 + 32} = \frac{180}{212} = 0.8491 = 84.91\%$$

Specificity (Non-appendicitis)

$$\frac{TN}{TN + FP} = \frac{170}{170 + 42} = \frac{170}{212} = 0.8019 = 80.19\%$$

e. F1-Score (appendicitis)

$$F1 = 2 \frac{(0.8416 \times 0.8019)}{0.8416 + 0.8019} = 0.8213 = 82.13\%$$

F1-Score (Non-appendicitis)

$$F1 = 2 \frac{(0.8108 \times 0.8491)}{0.8108 + 0.8491} = 0.8295 = 82.95\%$$

f. Marco Precision

$$\frac{0.8416 + 0.8108}{2} = 0.8262$$

Macro Recall

$$\frac{0.8019 + 0.8491}{2} = 0.8255$$

Macro F1

$$\frac{0.8213 + 0.8295}{2} = 0.8254$$

g. Weighted Precision

$$\frac{(0.8416 \times 212) + (0.8108 \times 212)}{424} = 0.8262$$

Weighted Recall

$$\frac{(0.8019 \times 212) + (0.8491 \times 212)}{424} = 0.8255$$

Weighted F1

$$\frac{(0.8213 \times 212) + (0.8295 \times 212)}{424} = 0.8254$$

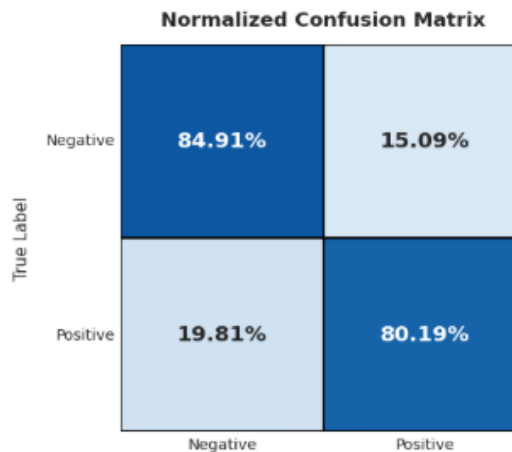
Untuk mengevaluasi performa model VGG-19 secara kuantitatif, digunakan metrik precision, recall, F1-score, dan accuracy berdasarkan hasil klasifikasi pada data uji. Hasil evaluasi tersebut ditunjukkan pada Gambar 6

Class	Precision	Recall	F1-Score	Support
Negative	0.8108	0.8491	0.8295	212
Positive	0.8416	0.8019	0.8213	212
Accuracy			0.8255	424
Macro Avg	0.8262	0.8255	0.8254	424
Weighted Avg	0.8262	0.8255	0.8254	424

Gambar 6. Hasil Model confusion matrix data validation

Berdasarkan Gambar 6 Untuk mengevaluasi performa model VGG-19 secara kuantitatif, digunakan metrik evaluasi berupa precision, recall, F1-score, dan accuracy yang dihitung berdasarkan hasil klasifikasi pada data uji. Hasil evaluasi tersebut ditunjukkan pada Gambar 3.X. Berdasarkan gambar tersebut, model memperoleh nilai accuracy sebesar 0,8255, yang menunjukkan bahwa model mampu mengklasifikasikan 82,55% dari total 424 citra dengan benar. Pada kelas non-appendicitis (negative), model mencapai precision sebesar 0,8108, recall sebesar 0,8491, dan F1-score sebesar 0,8295, yang menunjukkan bahwa model cukup baik dalam mengenali citra non-appendicitis secara konsisten.

Untuk mengevaluasi performa klasifikasi model VGG-19, digunakan normalized confusion matrix hasil klasifikasi dalam bentuk persentase untuk setiap kelas. Visualisasi ini bertujuan untuk menunjukkan kemampuan model dalam membedakan antara kelas appendicitis dan non-appendicitis berdasarkan proporsi prediksi yang benar dan salah. Hasil normalized confusion matrix model VGG-19 ditunjukkan pada Gambar 7.

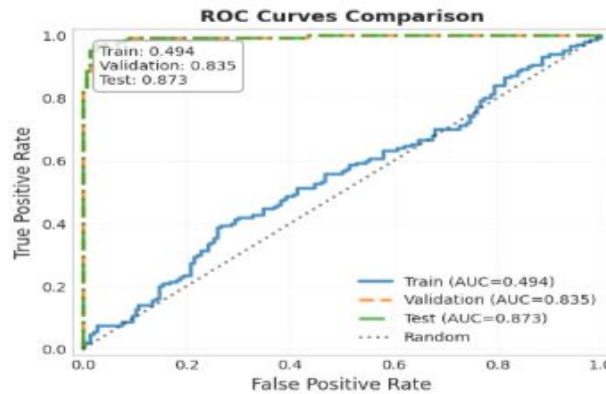


Gambar 7. Normalize matrix vgg-19

Berdasarkan Gambar 7. Sejalan dengan confusion matriks normalized confusion matrix menunjukkan bahwa model VGG-19 hanya mampu mengenali kelas appendicitis dengan akurasi 80,19%, sementara 19,81% kasusnya salah diklasifikasikan sebagai non-appendicitis. Sebaliknya, kelas non-appendicitis dikenali hampir sempurna dengan akurasi

84,91%. Dan hanya 15,09% kasus yang mengalami kesalahan klasifikasi. Hasil ini menegaskan bahwa model sangat bias terhadap kelas non-appendicitis dan gagal membedakan kasus appendicitis secara efektif.

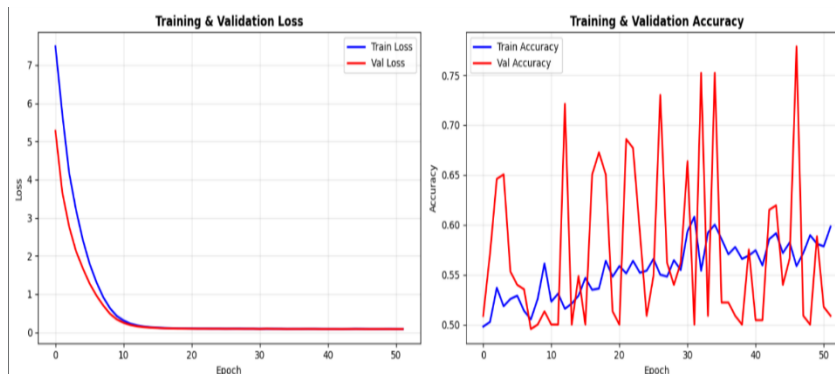
Model VGG-19 dalam membedakan antara kelas appendicitis dan non-appendicitis, digunakan kurva Receiver Operating Characteristic (ROC). Kurva ROC menunjukkan hubungan antara True Positive Rate (TPR) dan False Positive Rate (FPR) pada berbagai nilai threshold, serta memberikan gambaran mengenai kemampuan model dalam melakukan klasifikasi secara keseluruhan. Hasil kurva ROC model VGG-19 pada data training, validation, dan testing ditunjukkan pada Gambar 8.



Gambar 8. Roc curve vgg-19

Kurva ROC pada model VGG-19 menunjukkan nilai AUC pada data latih sebesar 0,494, nilai pada validasi dan pengujian mencapai 0,835 dan 0,873 yang mengindikasikan bahwa secara teori model memiliki kemampuan diskriminasi yang cukup baik dalam membedakan kelas appendicitis dan non-appendicitis. Meskipun demikian, performa ini tidak terlihat pada hasil prediksi aktual, Rendahnya nilai AUC ini sejalan dengan hasil evaluasi sebelumnya, dimana performa klasifikasi vgg-19 menunjukkan banyak kesalahan prediksi.

Grafik ini memberikan gambaran mengenai bagaimana model belajar dari data pelatihan serta bagaimana performanya ketika diuji pada data validasi yang tidak digunakan dalam proses pelatihan. Hasil training model VGG-19 ditunjukkan pada Gambar 9

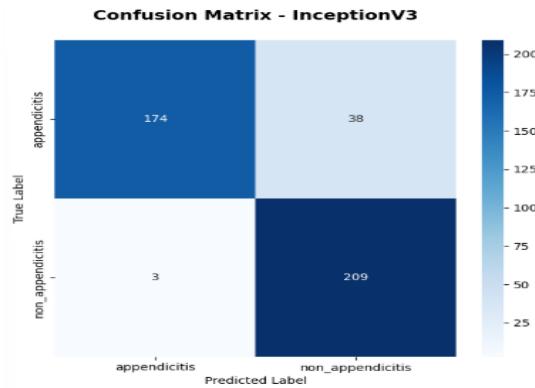


Gambar 9. Training history vgg-19

Grafik training history pada model VGG-19 menunjukkan adanya ketidakseimbangan yang jelas antara kinerja data pelatihan dan data validasi. Pada grafik loss, terlihat bahwa training loss dan validation loss sama-sama menurun tajam di awal epoch dan kemudian konvergen mendekati nilai yang sangat kecil, yang menandakan proses optimasi berjalan dengan baik. Namun, pada grafik akurasi, akurasi data pelatihan meningkat secara bertahap dari sekitar 50% hingga mendekati 60%, sementara akurasi validasi berfluktuasi secara tajam di kisaran 50–75% tanpa menunjukkan tren peningkatan yang stabil. Ketidakkonsistenan akurasi validasi ini mengindikasikan bahwa meskipun model mampu mempelajari pola pada data pelatihan, performanya pada data yang tidak dilatih masih belum stabil

3.6 Evaluasi Model Inceptionv3

InceptionV3 menunjukkan kinerja yang stabil dan mampu melakukan generalisasi dengan baik. Pada tahap training awal, akurasi meningkat hingga sekitar 77-83%, disertai validation loss yang menurun dan akurasi validasi yang berkisar antara 55%–69%. Pada tahap fine-tuning, performa model meningkat signifikan dengan akurasi pelatihan mencapai 95-97% dan akurasi validasi yang konsisten tinggi pada rentang 72%–77%. Pola ini menunjukkan bahwa InceptionV3 tidak hanya mampu mempelajari data pelatihan dengan baik, tetapi juga mempertahankan performa yang konsisten pada data validasi, sehingga mencerminkan kemampuan generalisasi yang baik.



Gambar 10. Confusion matrix inception V3

Confusion matrix menunjukkan bahwa model InceptionV3 mampu mengklasifikasikan kasus appendicitis dan non-appendicitis dengan performa yang cukup baik dan lebih seimbang. Berdasarkan gambar, pada kelas appendicitis, model berhasil memprediksi 174 kasus dengan benar, sementara 38 kasus masih salah diklasifikasikan sebagai non-appendicitis. Pada kelas non-appendicitis, InceptionV3 menunjukkan kinerja yang sangat baik dengan 209 kasus berhasil diprediksi secara benar dan hanya 3 kasus yang keliru diklasifikasikan sebagai appendicitis. Distribusi kesalahan yang relatif kecil, khususnya pada kelas non-appendicitis, menunjukkan bahwa model ini memiliki kemampuan diskriminasi yang kuat dalam membedakan kedua kelas dan secara keseluruhan memperlihatkan performa klasifikasi yang lebih stabil dan seimbang dibandingkan model VGG-19

Berikut ini perhitungan matrix pada data validasi dalam klasifikasi penyakit Appendicitis sesuai kelasnya:

a. Accuracy (appendicitis)

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} + \frac{174 + 209}{424} = \frac{383}{424} = 0.9033 = 90.33\%$$

Accuracy (Non-Appendicitis)

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} + \frac{174 + 209}{424} = \frac{383}{424} = 0.9033 = 90.33\%$$

b. Precision(appendicitis)

$$precision = \frac{TP}{TP + FP} + \frac{174}{174 + 11} = \frac{174}{188} = 0.9831 = 98.31\%$$

precision (Non-appendicitis)

$$\frac{TP}{TP + FP} + \frac{209}{209 + 38} = \frac{209}{274} = 0.8462 = 84.62\%$$

c. Sensitivity/recall (appendicitis)

$$\frac{TP}{TP + FN} + \frac{174}{174 + 38} = \frac{174}{212} = 0.8208 = 82.08\%$$

Sensitivity/recall (Non-appendicitis)

$$\frac{TP}{TP + FN} + \frac{209}{209 + 3} = \frac{209}{212} = 0.9858 = 98.58\%$$

d. Specificity (appendicitis)

$$\frac{TN}{TN + FP} + \frac{209}{209 + 3} = \frac{209}{212} = 0.9858 = 98.58\%$$

Specificity (Non-appendicitis)

$$\frac{TN}{TN + FP} + \frac{174}{174 + 38} = \frac{174}{212} = 0.8208 = 82.08\%$$

e. F1-Score (appendicitis)

$$F1 = 2 \frac{(0.9831 \times 0.8208)}{0.9831 + 0.8208} = 0.8946 = 89.46\%$$

F1-Score (Non-appendicitis)

$$F1 = 2 \frac{0.8462 \times 0.9858}{0.8462 + 0.9858} = 0.9107 = 91.07\%$$

f. Marco Precision

$$\frac{0.9831 + 0.8462}{2} = 0.9146$$

Macro Recall

$$\frac{0.8208 + 0.9858}{2} = 0.9033$$

Macro F1

$$\frac{0.8946 + 0.9107}{2} = 0.9026$$

g. Weighted Precision

$$\frac{(0.9831 \times 212) + (0.8462 \times 212)}{424} = 0.9146$$

Weighted Recall

$$\frac{(0.8208 \times 212) + (0.9858 \times 212)}{424} = 0.9033$$

Weighted F1

$$\frac{(0.8946 \times 212) + (0.9107 \times 212)}{424} = 0.9026$$

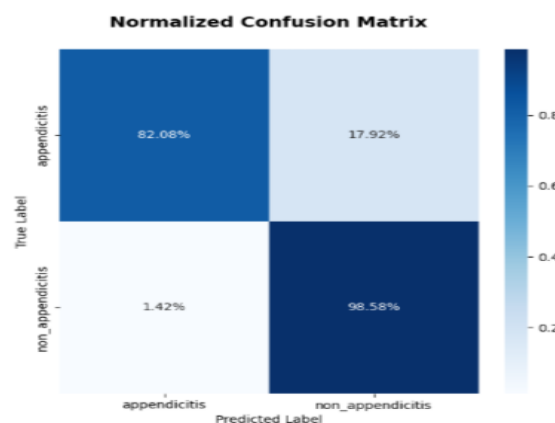
Untuk mengevaluasi performa model InceptionV3 secara kuantitatif, digunakan metrik precision, recall, F1-score, dan accuracy berdasarkan hasil klasifikasi pada data uji. Hasil evaluasi tersebut ditunjukkan pada Gambar 3.11

DETAILED CLASSIFICATION REPORT (Test Set):				
	precision	recall	f1-score	support
appendicitis	0.9831	0.8208	0.8946	212
non_appendicitis	0.8462	0.9858	0.9107	212
accuracy			0.9033	424
macro avg	0.9146	0.9033	0.9026	424
weighted avg	0.9146	0.9033	0.9026	424

Gambar 11. Hasil Model confusion matrix data validation

Berdasarkan gambar 11 model InceptionV3 memperoleh nilai accuracy sebesar 0,9033, yang menunjukkan bahwa model mampu mengklasifikasikan 90,33% dari total 424 citra dengan benar. Pada kelas appendicitis, model mencapai precision sebesar 0,9831, recall sebesar 0,8208, dan F1-score sebesar 0,8946. Nilai precision yang sangat tinggi menunjukkan bahwa model memiliki tingkat ketepatan yang sangat baik dalam memprediksi kasus appendicitis, dengan jumlah kesalahan prediksi positif yang sangat rendah. Sementara itu, pada kelas non-appendicitis, model memperoleh precision sebesar 0,8462, recall sebesar 0,9858, dan F1-score sebesar 0,9107, yang menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam mengenali hampir seluruh citra non-appendicitis secara benar.

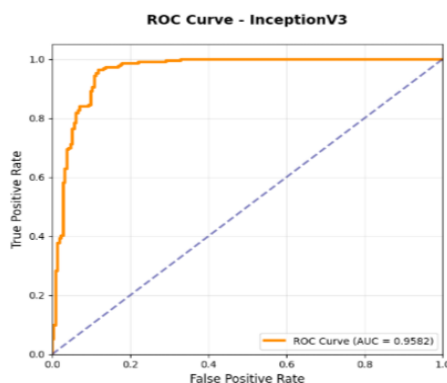
Visualisasi ini bertujuan untuk menunjukkan kemampuan model dalam mengenali dan membedakan antara kelas appendicitis dan non-appendicitis berdasarkan proporsi prediksi yang benar dan salah. Hasil normalized confusion matrix model InceptionV3 ditunjukkan pada Gambar 12



Gambar 12. Normalize matrix inception V3

Normalized confusion matrix memperlihatkan bahwa InceptionV3 mengenali 82,08% kasus appendicitis secara benar dan 17,92% kasus non-appendicitis masih salah diklasifikasi. Sedangkan kelas non-appendicitis model menunjukkan performa yang sangat tinggi dengan tingkat pengenalan benar sebesar 98,58%, dan hanya 1,42% kasus yang mengalami kesalahan klasifikasi. Tingkat kesalahan yang relatif rendah ini menunjukkan bahwa InceptionV3 memiliki kemampuan klasifikasi yang sangat baik dan stabil, khususnya dalam membedakan kasus non-appendicitis, serta lebih andal untuk digunakan dalam diagnosis medis.

Kurva ROC menunjukkan hubungan antara True Positive Rate (TPR) dan False Positive Rate (FPR) pada berbagai nilai threshold, serta digunakan untuk mengukur kemampuan diskriminasi model secara keseluruhan melalui nilai Area Under Curve (AUC). Kurva ROC model InceptionV3 ditunjukkan pada Gambar 3.13



Gambar 13. Roc curve inception V3

Kurva ROC menghasilkan nilai AUC sebesar 0.9502, yang mengindikasikan bahwa model memiliki kemampuan diskriminasi yang sangat baik dalam membedakan kasus appendicitis dan non-appendicitis.

Grafik ini memberikan gambaran mengenai bagaimana model mempelajari pola dari data pelatihan serta bagaimana performanya pada data validasi yang digunakan untuk mengevaluasi kemampuan generalisasi model. Hasil training model InceptionV3 ditunjukkan pada Gambar 14



Gambar 14. Training history inception-V3

Grafik training history pada model InceptionV3 menunjukkan bahwa akurasi pelatihan meningkat secara bertahap dan relatif stabil, dimulai dari kisaran 75–85% pada fase awal (phase 1) dan kemudian meningkat tajam hingga melampaui 90% setelah tahap fine-tuning (phase 2). Sementara itu, akurasi validasi cenderung berfluktuasi, namun tetap berada pada rentang yang cukup baik, yaitu sekitar 60–80%. Pada grafik loss, terlihat bahwa training loss menurun secara konsisten sepanjang proses pelatihan, terutama setelah fine-tuning, sedangkan validation loss mengikuti pola yang relatif stabil meskipun masih mengalami fluktuasi kecil. Pola ini menunjukkan bahwa model InceptionV3 mampu mempelajari fitur data secara efektif dan memiliki kemampuan generalisasi yang cukup baik, sehingga performa model tergolong stabil dan lebih optimal dibandingkan model sebelumnya.

3.7 Hasil Perbandingan

a. Perbandingan kelas appendicitis

Untuk membandingkan performa model VGG-19 dan InceptionV3 dalam mengklasifikasikan citra appendicitis, digunakan beberapa metrik evaluasi yaitu accuracy, precision, recall, dan F1-score. Metrik ini memberikan gambaran mengenai tingkat ketepatan, kemampuan deteksi kasus positif, serta keseimbangan performa model dalam klasifikasi. Hasil perbandingan performa kedua model pada kelas appendicitis ditunjukkan pada Tabel 2.

Tabel 2. Tabel perbandingan kelas appendicitis

Model	Accuracy	Precision	Recall	F1-Score	Support
VGG-19	0.8255	0.8416	0.8019	0.8213	212
Inceptionv3	0.9033	0.9831	0.8208	0.8946	212

Hasil pada Tabel 2 menunjukkan adanya perbedaan performa yang cukup signifikan antara model VGG-19 dan InceptionV3 dalam mengklasifikasikan kasus appendicitis. Berdasarkan metrik evaluasi, model VGG-19 memperoleh nilai accuracy sebesar 0,8255, precision 0,8416, recall 0,8019, dan F1-score 0,8213 dengan jumlah data uji (support) sebanyak 212, yang mengindikasikan bahwa kemampuan model ini dalam memberikan prediksi yang konsisten dan seimbang antara ketepatan prediksi positif dan deteksi seluruh kasus appendicitis masih tergolong terbatas. Sebaliknya, InceptionV3 menunjukkan performa yang lebih unggul dengan accuracy 0,9033, precision 0,9831, recall 0,8208, dan

F1-score 0,8946 pada jumlah data yang sama, yang menandakan tingkat ketepatan prediksi positif yang sangat tinggi serta kemampuan klasifikasi yang lebih stabil. Berdasarkan perbandingan tersebut, dapat disimpulkan bahwa InceptionV3 lebih layak digunakan dibandingkan VGG-19 dalam sistem klasifikasi appendicitis karena memberikan performa yang lebih akurat.

b. Perbandingan kelas non-appendicitis

Untuk membandingkan performa model VGG-19 dan InceptionV3 dalam mengklasifikasikan citra appendicitis, digunakan beberapa metrik evaluasi yaitu accuracy, precision, recall, dan F1-score. Metrik ini memberikan gambaran mengenai tingkat ketepatan, kemampuan deteksi kasus positif, serta keseimbangan performa model dalam klasifikasi. Hasil perbandingan performa kedua model pada kelas appendicitis ditunjukkan pada Tabel 3.

Tabel 3. Tabel perbandingan kelas non-appendicitis

Model	Accuracy	Precision	Recall	F1-Score	Support
VGG-19	0.8255	0.8108	0.8491	0.8295	212
Inceptionv3	0.9033	0.8462	0.9858	0.9107	212

Hasil pada Tabel 3.3 menunjukkan perbandingan performa model VGG-19 dan InceptionV3 dalam mengklasifikasikan kelas non-appendicitis. Model VGG-19 memperoleh nilai accuracy sebesar 0,8255, precision 0,8108, recall 0,8491, dan F1-score 0,8295 dengan jumlah data uji (support) sebanyak 212, yang menunjukkan bahwa meskipun model ini cukup baik dalam mengenali kasus non-appendicitis, tingkat ketepatan dan konsistensinya masih belum optimal. Sebaliknya, InceptionV3 menunjukkan performa yang lebih unggul dengan accuracy 0,9033, precision 0,8462, recall 0,9858, dan F1-score 0,9107 pada jumlah data yang sama, yang mengindikasikan kemampuan yang sangat baik dalam mendeteksi hampir seluruh kasus non-appendicitis secara benar. Perbedaan hasil ini menegaskan bahwa InceptionV3 memiliki performa klasifikasi yang lebih stabil dan akurat dibandingkan VGG-19.

c. Perbandingan kinerja model global

Untuk mengevaluasi performa keseluruhan model VGG-19 dan InceptionV3, dilakukan perbandingan berdasarkan beberapa parameter penting, yaitu accuracy dan loss pada tahap training, validasi, dan testing, serta waktu pelatihan dan jumlah epoch yang digunakan. Parameter-parameter ini memberikan gambaran mengenai kemampuan model dalam mempelajari data, melakukan generalisasi terhadap data baru, serta efisiensi komputasi selama proses pelatihan. Hasil perbandingan kinerja global kedua model ditunjukkan pada Tabel 4.

Tabel 4. Tabel perbandingan global

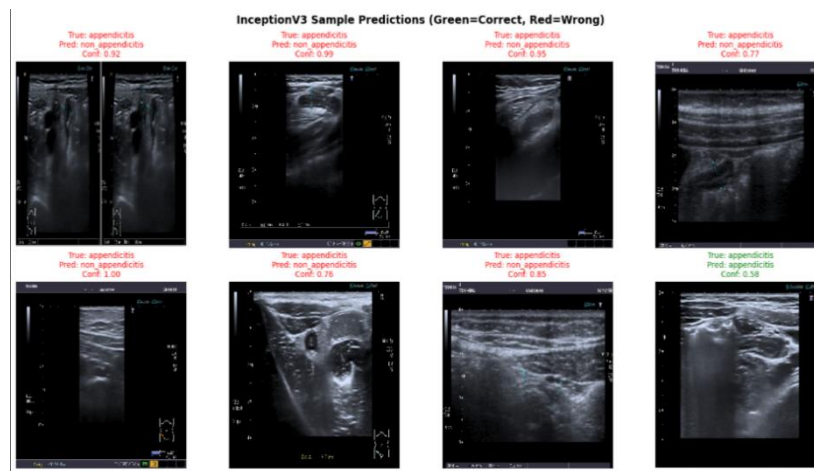
	VGG-19	Inceptionv3
Accuracy Training	0.4941	0.4888
Loss Training	0.0873	0.2491
Accuracy Validasi	0.7788	0.8982
Loss Validasi	0.0819	0.3035
Accuracy Testing	0.8255	0.9033
Loss Testing	0.0817	0.2992
Time	36.21 Menit	51 Menit
Epoch	50 Epoch	50 Epoch

Berdasarkan tabel perbandingan kinerja global antara VGG-19 dan InceptionV3 menunjukkan perbedaan yang sangat signifikan pada tahap training, validasi, dan pengujian. Model VGG-19 memperoleh akurasi training sebesar 0.4941 dengan loss 0.0873, akurasi validasi 0.7788 dengan loss 0.0819, serta akurasi testing 0.8255 dengan loss 0.0817, yang mengindikasikan bahwa performa model ini relatif rendah dan cenderung stagnan pada seluruh tahap evaluasi. Sebaliknya, InceptionV3 menunjukkan kinerja yang jauh lebih unggul training 0.4888 dengan loss training yang sangat rendah sebesar 0.2491, akurasi validasi yang tinggi yaitu 0.8982 dengan loss 0.3035, serta mampu mempertahankan performa yang sangat baik pada tahap testing dengan akurasi 0.9033 dan loss 0.2992. Pola ini menunjukkan bahwa InceptionV3 tidak hanya mampu mempelajari data dengan baik, tetapi juga memiliki kemampuan generalisasi yang kuat terhadap databaru, sehingga secara keseluruhan model ini lebih stabil, akurat, dan lebih layak digunakan sebagai sistem klasifikasi appendicitis dibandingkan dengan VGG-19. Meskipun InceptionV3 membutuhkan waktu pelatihan yang lebih lama, yaitu 51 menit dibandingkan 36,21 menit pada VGG-19 dengan jumlah epoch yang sama, pola hasil ini menunjukkan bahwa InceptionV3 memiliki kemampuan generalisasi yang lebih kuat terhadap data baru, sehingga secara keseluruhan lebih stabil, lebih akurat, dan lebih layak digunakan sebagai sistem klasifikasi appendicitis dibandingkan model VGG-19.

d. Perbandingan hasil prediksi klasifikasi

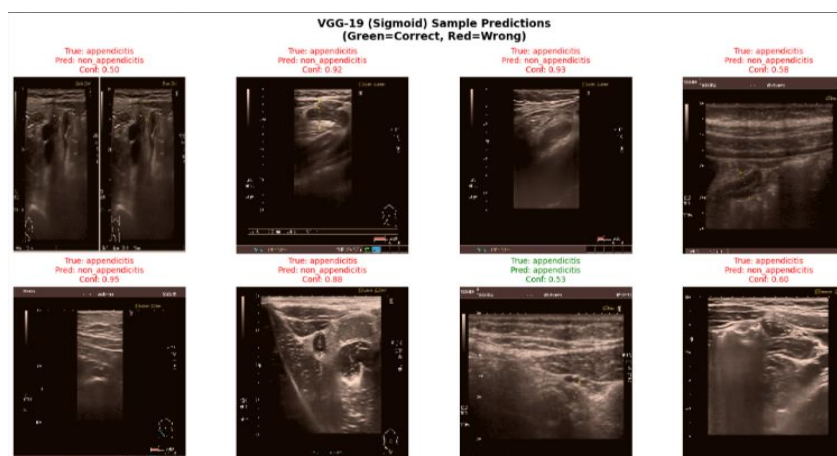
Untuk memberikan gambaran visual mengenai hasil klasifikasi model VGG-19, dilakukan pengujian terhadap beberapa sampel citra ultrasonografi dari data uji. Visualisasi ini bertujuan untuk menunjukkan bagaimana model memprediksi kelas citra serta tingkat kepercayaan (confidence) terhadap prediksi tersebut. Selain itu, visualisasi ini juga

membantu dalam mengidentifikasi kesalahan klasifikasi yang dilakukan oleh model. Hasil prediksi model VGG-19 terhadap beberapa sampel citra ditunjukkan pada Gambar 15



Gambar 15. Gambar sample hasil klasifikasi VGG-19

Untuk memberikan gambaran visual mengenai hasil klasifikasi model InceptionV3, dilakukan pengujian terhadap beberapa sampel citra ultrasonografi dari data uji. Visualisasi ini bertujuan untuk menunjukkan kemampuan model dalam memprediksi kelas citra serta tingkat kepercayaan (confidence) terhadap hasil prediksi. Selain itu, visualisasi ini juga digunakan untuk mengevaluasi konsistensi dan akurasi model dalam mengenali karakteristik citra appendicitis dan non-appendicitis. Hasil prediksi model InceptionV3 terhadap beberapa sampel citra ditunjukkan pada Gambar 3.16



Gambar 16. Gambar sample hasil klasifikasi Inceptionv3

Berdasarkan gambar Diatas Analisis visual terhadap sampel prediksi menunjukkan perbedaan kinerja yang sangat jelas antara VGG-19 dan InceptionV3. Pada VGG-19, sebagian besar citra appendicitis salah diklasifikasikan sebagai non-appendicitis, bahkan dengan tingkat keyakinan yang tinggi. Hal ini menegaskan bahwa model gagal mengenali pola-pola khas appendicitis pada ultrasonografi dan memiliki kalibrasi confidence yang buruk. Sebaliknya, InceptionV3 memperlihatkan prediksi yang lebih akurat dan konsisten, dimana mayoritas citra appendicitis berhasil diidentifikasi dengan benar disertai nilai confidence yang tinggi. Bahkan pada kasus visual yang lebih sulit, InceptionV3 tetap mampu menghasilkan prediksi yang mendekati benar dibandingkan VGG-19. Secara keseluruhan, hasil visual ini mendukung temuan kuantitatif bahwa InceptionV3 memiliki pemahaman fitur yang lebih baik, keandalan prediksi yang lebih tinggi, dan jauh lebih sesuai digunakan dalam konteks deteksi appendicitis dibandingkan VGG-19.

Secara lebih rinci, tahapan penerapan algoritma dimulai dari proses pre-processing citra (resize, normalisasi, dan augmentasi), dilanjutkan dengan pelatihan model menggunakan skema transfer learning pada kedua arsitektur, kemudian dilakukan pengujian menggunakan data uji terpisah untuk memperoleh metrik evaluasi berupa accuracy, precision, recall, F1-score, AUC, serta confusion matrix. Tahapan ini memastikan bahwa hasil yang diperoleh benar-benar merefleksikan kemampuan generalisasi model terhadap data baru.

Hasil pengujian menunjukkan bahwa VGG-19 menghasilkan akurasi yang relatif rendah serta nilai recall kelas appendicitis yang sangat kecil, yang berarti model memiliki tingkat false negative yang tinggi. Kondisi ini berisiko dalam konteks klinis karena dapat menyebabkan kasus appendicitis tidak terdeteksi (missed diagnosis). Sebaliknya, InceptionV3 menunjukkan performa yang jauh lebih seimbang antara sensitivity (recall) dan specificity, dengan nilai

F1-score dan AUC yang tinggi, yang mengindikasikan kemampuan klasifikasi yang stabil dan diskriminatif terhadap kedua kelas.

Jika dibandingkan dengan penelitian sejenis yang berfokus pada segmentasi organ menggunakan U-Net atau klasifikasi penyakit lain berbasis CNN, hasil penelitian ini menunjukkan bahwa pendekatan klasifikasi langsung menggunakan arsitektur InceptionV3 lebih efektif untuk tujuan identifikasi cepat kondisi appendicitis. Hal ini memperkuat temuan bahwa arsitektur dengan modul multi-skala seperti Inception mampu menangkap variasi pola tekstur pada citra USG yang cenderung memiliki noise dan kontras rendah.

Dengan demikian, pembahasan ini menegaskan bahwa perbedaan arsitektur berpengaruh signifikan terhadap kemampuan ekstraksi fitur dan performa klasifikasi. InceptionV3 tidak hanya unggul secara kuantitatif, tetapi juga secara visual dan klinis lebih layak dipertimbangkan sebagai dasar pengembangan sistem pendukung keputusan untuk deteksi appendicitis berbasis citra medis.

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa model InceptionV3 memiliki performa yang jauh lebih unggul dan stabil dibandingkan model VGG-19 dalam mengklasifikasikan penyakit *appendicitis*. Hal ini dibuktikan melalui capaian akurasi pengujian (*testing accuracy*) sebesar 90,33% pada InceptionV3, yang secara signifikan melampaui VGG-19 yang hanya mencapai 82,55%. Model VGG-19 menunjukkan indikasi *overfitting* yang cukup kuat, di mana model sulit melakukan generalisasi pada data baru yang ditandai dengan fluktuasi akurasi validasi yang tajam serta ketidakmampuan mengenali pola khas pada citra ultrasonografi secara konsisten, sehingga kurang ideal untuk digunakan dalam konteks medis. Keunggulan InceptionV3 juga tercermin dari kemampuan diskriminasi kelas yang sangat baik dengan nilai AUC mencapai 0,9502, yang menunjukkan tingkat keandalan tinggi dalam membedakan kasus *appendicitis* dan *non-appendicitis*. Secara visual, InceptionV3 mampu memberikan prediksi yang lebih akurat dengan nilai kepercayaan (*confidence*) yang tinggi, bahkan pada sampel citra yang memiliki tingkat kesulitan visual lebih besar. Oleh karena itu, meskipun membutuhkan waktu pelatihan yang sedikit lebih lama yaitu 51 menit dibandingkan VGG-19 yang hanya 36,21 menit, InceptionV3 tetap menjadi pilihan yang jauh lebih layak dan direkomendasikan sebagai sistem klasifikasi diagnosis *appendicitis* karena stabilitas dan akurasi menyeluruh yang dihasilkannya.

REFERENCES

- [1] Abidah, S. H. Wijoyo, dan K. Rahman, "Pengaruh Platform Visual Studio Code Terhadap Hasil Belajar Siswa pada Mata Pelajaran Pemrograman Dasar Kelas X Jurusan Teknik Komputer dan Jaringan SMKN 3 Malang," *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, vol. 9, no. 3, pp. 2548–2964, 2025. [Online]. Available: <http://j-ptiik.ub.ac.id>
- [2] P. Arafin, A. Issa, dan A. H. M. M. Billah, "Performance Comparison of Multiple Convolutional Neural Networks for Concrete Defects Classification," *Sensors*, vol. 22, no. 22, 2022. doi: 10.3390/s22228714
- [3] T. Arif dan R. S. I. Tirtana, "Implementasi Convolutional Neural Network dengan Arsitektur," vol. 8, no. 1, pp. 41–47, 2024.
- [4] Zufria, I., Harumy, T. H. F., & Efendi, S. (2025). Identifying the best models for BISINDO alphabet gesture classification to support the communication needs of the deaf community. *Eastern-European Journal of Enterprise Technologies*, 6(2(138)), 26–41. <https://doi.org/10.15587/1729-4061.2025.332096>
- [5] B. T. Baştuğ, G. Güneri, M. S. Yıldırım, K. Çorbacı, dan E. Dandıl, "Fully Automated Detection of the Appendix Using U-Net Deep Learning Architecture in CT Scans," *Journal of Clinical Medicine*, vol. 13, no. 19, 2024. doi: 10.3390/jcm13195893
- [6] O. Bobeică dan D. Iorga, "Artificial neural networks development in prosthodontics - a systematic mapping review," *Journal of Dentistry*, vol. 151, Feb. 2024. doi: 10.1016/j.jdent.2024.105385
- [7] S. Dwianto, F. N. Mubarak, D. Satriatama, T. Agustin, dan S. Indonesia, "Penerapan Convolutional Neural Network (CNN) dalam ...," Nov. 2024, pp. 270–280.
- [8] H. Han et al., "Trends and levels of the global, regional, and national burden of appendicitis between 1990 and 2021: findings from the Global Burden of Disease Study 2021," *The Lancet Gastroenterology and Hepatology*, vol. 9, no. 9, pp. 825–858, 2024. doi: 10.1016/S2468-1253(24)00157-2
- [9] A. R. Hidayat dan V. Lusiana, "Deteksi Jenis Sayuran dengan Tensorflow Dengan Metode Convolutional Neural Network," *Jurnal Sains Komputer & Informatika (J-SAKTI)*, vol. 6, no. 2, pp. 1032–1040, 2022.
- [10] M. Issaiy, D. Zarei, dan A. Saghazadeh, "Artificial Intelligence and Acute Appendicitis: A Systematic Review of Diagnostic and Prognostic Models," *World Journal of Emergency Surgery*, vol. 18, no. 1, pp. 1–31, 2023. doi: 10.1186/s13017-023-00527-2
- [11] A. W. Kosman, Y. Wahyuningsih, dan F. Mahendrasusila, "Pengujian Metode Inception V3 dalam Mengidentifikasi Penyakit Kanker Kulit ...," vol. 10, no. 1, pp. 136–146, 2024.
- [12] D. Marcella, Y. Yohannes, dan S. Devella, "Klasifikasi Penyakit Mata Menggunakan Convolutional Neural Network Dengan Arsitektur VGG-19," *Jurnal Algoritma*, vol. 3, no. 1, pp. 60–70, 2022. doi: 10.35957/algoritma.v3i1.3331
- [13] P. Palupiningsih, A. R. Sujiwanto, dan R. R. B. Prawirodirjo, "Analisis Perbandingan Performa Model Klasifikasi Kesehatan Daun Tomat menggunakan arsitektur VGG, MobileNet, dan Inception V3," *Jurnal Ilmu Komputer Dan Agri-Informatika*, vol. 10, no. 1, pp. 98–110, 2023. doi: 10.29244/jika.10.1.98-110
- [14] F. Parajun, R. N. Aziza, dan D. Kuswardani, "Implementasi Algoritma Convolutional Neural Network Dalam Mengklasifikasi Kesegaran Buah Berdasarkan Citra Buah," *Kilat*, vol. 11, no. 1, pp. 1–9, 2022. doi: 10.33322/kilat.v10i2.1458
- [15] S. H. Park et al., "Comparison between single and serial computed tomography images in classification of acute appendicitis, acute right-sided diverticulitis, and normal appendix using EfficientNet," *PLoS ONE*, vol. 18, no. 5, May 2023. doi: 10.1371/journal.pone.0281498

- [16] Y. Pratama, E. Rasywir, F. Fachruddin, D. Kisbianty, dan B. Irawan, "Eksperimen Layer Pooling menggunakan Standar Deviasi untuk Klasifikasi Dataset Citra Wajah dengan Metode CNN," *Building of Informatics, Technology and Science (BITS)*, vol. 5, no. 1, pp. 200–210, 2023. doi: 10.47065/bits.v5i1.3604
- [17] R. Purnamasari, F. I. Syahrudin, A. M. Dirgahayu, D. Iskandar, dan F. Fadhila, "Karakteristik Klinis Penderita Apendisitis," *UMI Medical Journal*, vol. 8, no. 2, pp. 117–126, 2023. doi: 10.33096/umj.v8i2.241
- [18] P. Rathna Sekhar dan S. Goud, "Collaborative Learning Techniques in Python Programming: A Case Study with CSE Students at Anurag University," *Journal of Engineering Education Transformations*, vol. 38, Special Issue 1, pp. 243–249, 2024. doi: 10.16920/jeet/2024/v38is1.24238
- [19] S. Sriani dan A. Nabila, "Implementasi Deep Learning Untuk Mengidentifikasi Umur Manusia Menggunakan Convolutional Neural Network (CNN)," *Jurnal Informatika Dan Teknik Elektro Terapan*, vol. 12, no. 3, pp. 1836–1843, 2024. doi: 10.23960/jitet.v12i3.4457
- [20] J. Strohäker, M. Brüscke, Y. S. Feng, C. Beltzer, A. Königsrainer, dan R. Ladurner, "Predicting complicated appendicitis is possible without the use of sectional imaging—presenting the NoCtApp score," *International Journal of Colorectal Disease*, vol. 38, no. 1, pp. 1–10, 2023. doi: 10.1007/s00384-023-04501-x
- [21] S. Supiyandi, T. Panggabean, N. Ramadhan, S. R. Dewi, dan S. Yusra, "Deteksi Tepi Sederhana Pada Citra Menggunakan Operator Sobel," *Repeater: Publikasi Teknik Informatika Dan Jaringan*, vol. 2, no. 3, pp. 43–56, 2024. doi: 10.62951/repeater.v2i3.90
- [22] N. J. Telambanua, N. Nofriadi, dan A. Dermawan, "Sistem Pakar Untuk Mendeteksi Penyakit Mata Menerapkan Metode Case Based Reasoning," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 2, pp. 570–580, 2022. doi: 10.47065/bits.v4i2.2116
- [23] E. Yuninsar, E. Hasmin, I. Samsie, dan T. Indah, "Analisis Perbandingan Metode Convolutional Neural Network (CNN) dan Artificial Neural Network (ANN)," vol. 03, pp. 321–333, 2024.
- [24] Y. Zheng et al., "MAIT cell activation and recruitment in inflammation and tissue damage in acute appendicitis," 2024. doi: 10.1126/sciadv.adn6331
- [25] Z. Zheng, Y. Xiang, Y. Qi, Y. Lin, dan H. Zhang, "Fully Convolutional Neural Networks for High-Precision Medical Image Analysis," 2024.
- [26] I. Zufria, Y. Farentika, U. Islam, dan N. Sumatera, "Penerapan Metode TOPSIS Dan SAW Dalam Menentukan Penerimaan Anggota OSIS Baru," vol. 5, no. 1, pp. 154–166, 2023.