Jurnal Teliska
Volume 19 No.3 November 2026
Doi: 10.36257/teliska.v19i3.12682

P-ISSN: 2085-0786
E-ISSN: 2654-2765

# ANALISIS PERFORMA MODEL CNN MELALUI EVALUASI EKSTERNAL PADA SISTEM KLASIFIKASI CITRA MAMMOGRAFI BERBASIS WEBSITE LOKAL

**Suroso[1]*, Sopian Soim[2], Cik Sadan[3], Nurul Mardhiyah[4], Aisyah Salsabila Azzahrah[5], Adithya Putra[6]**
[1,2,4-6] Sarjana Terapan Teknik Telekomunikasi, Politeknik Negeri Sriwijaya, Palembang, Indonesia
[3] DIII Teknik Telekomunikasi, Politeknik Negeri Sriwijaya, Palembang, Indonesia
*Email korespondensi: suroso.0719@gmail.com

**Submitted** : 12 Oktober 2026 | **Accepted** : 2 November 2026 | **Published** : 11 November 2026

**ABSTRAK**

Kinerja model *Convolutional Neural Network* (CNN) dalam klasifikasi citra mammografi perlu dievaluasi menggunakan data dari sumber berbeda untuk mengetahui kemampuan generalisasi model. Penelitian ini bertujuan menganalisis performa model CNN EfficientNetB2 melalui evaluasi eksternal pada sistem klasifikasi citra mammografi berbasis website lokal. Pengujian dilakukan menggunakan 15 citra mammografi klinis sebagai data eksternal yang tidak dilibatkan dalam proses pengembangan model. Setiap citra diproses melalui sistem untuk memperoleh hasil klasifikasi, kemudian prediksi dibandingkan dengan label diagnosis pada data asli. Evaluasi kinerja dilakukan menggunakan *confusion matrix* dengan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Pada tahap pengembangan, model EfficientNetB2 menghasilkan *accuracy* sebesar 72%, *precision* sebesar 67%, *recall* sebesar 96%, dan *F1-score* sebesar 79%. Sementara itu, hasil evaluasi eksternal menunjukkan bahwa 13 dari 15 citra berhasil diklasifikasikan sesuai dengan label diagnosis, sedangkan 2 citra lainnya menghasilkan prediksi yang tidak sesuai. Berdasarkan *confusion matrix*, diperoleh *accuracy* sebesar 86,67%, *precision* sebesar 80%, *recall* sebesar 100%, dan *F1-score* sebesar 88,89%. Hasil penelitian menunjukkan bahwa model EfficientNetB2 memperoleh tingkat kesesuaian prediksi sebesar 86,67% pada pengujian menggunakan 15 citra mammografi eksternal melalui sistem berbasis website lokal. Hasil ini memberikan gambaran awal mengenai performa model pada data klinis dari sumber eksternal.

**Kata kunci:** *Convolutional Neural Network*; *EfficientNetB2*; Evaluasi Eksternal; Klasifikasi Citra Mammografi; Website Lokal.

**ABSTRACT**

*The performance of a Convolutional Neural Network (CNN) model for mammographic image classification needs to be evaluated using data from different sources to assess its generalization capability. This study aims to analyze the performance of the EfficientNetB2 CNN model through external evaluation using a local web-based mammographic image classification system. The evaluation was conducted using 15 clinical mammographic images as external test data that had not been used during model development. Each image was processed by the system to generate a classification prediction, which was subsequently compared with its corresponding original diagnostic label. Model performance was assessed using a confusion matrix and four evaluation metrics: accuracy, precision, recall, and F1-score. During the model development stage, EfficientNetB2 achieved an accuracy of 72%, a precision of 67%, a recall of 96%, and an F1-score of 79%. In the external evaluation, 13 out of 15 images were correctly classified, while the remaining two images were misclassified. Based on the confusion matrix, the external evaluation yielded an accuracy of 86.67%, a precision of 80%, a recall of 100%, and an F1-score of 88.89%. The results show that the EfficientNetB2 model achieved a prediction agreement rate of 86.67% when evaluated on 15 external mammographic images using a local web-based system. These findings provide a preliminary indication of the model's performance on clinical data from an external source.*

**Keywords:** *Convolutional Neural Network; EfficientNetB2; External Evaluation; Mammographic Image Classification; Web-Based Local System.*

## 1. PENDAHULUAN

Kanker merupakan salah satu penyakit tidak menular yang masih menjadi penyebab utama kematian di berbagai negara, penyakit ini ditandai oleh pertumbuhan sel yang tidak normal dan dapat menyebar ke jaringan maupun organ lain [1]. Berdasarkan laporan *Global Cancer Observatory* (GLOBOCAN) tahun

2022, kanker payudara merupakan jenis kanker dengan insidensi tertinggi pada perempuan di dunia [2]. Di Indonesia, kanker payudara juga menjadi jenis kanker dengan jumlah kasus tertinggi pada perempuan [3]. Tingginya angka kejadian tersebut menunjukkan pentingnya deteksi dini sebagai salah satu strategi penanganan kanker payudara karena dapat membantu mengidentifikasi kelainan pada tahap awal, sehingga berpotensi menurunkan angka kematian dan meningkatkan keberhasilan terapi [4]. Salah satu metode yang banyak digunakan dalam deteksi dini kanker payudara adalah mammografi. Pemeriksaan mammografi memanfaatkan sinar-X dosis rendah untuk menghasilkan citra jaringan payudara yang dapat digunakan untuk mengidentifikasi berbagai kelainan, seperti massa, mikrokalsifikasi, dan perubahan struktur jaringan [5]. Meskipun demikian, interpretasi citra mammografi masih bergantung pada pengalaman dan kemampuan radiolog sehingga berpotensi menimbulkan variasi pembacaan serta kesalahan diagnosis [6]. Karakteristik lesi kanker payudara yang memiliki kontras rendah dan pola visual kompleks juga dapat meningkatkan tingkat kesulitan dalam proses interpretasi citra [6]. Kondisi tersebut mendorong pemanfaatan *Artificial Intelligence* (AI), khususnya *Deep Learning* dengan pendekatan *Convolutional Neural Network* (CNN), sebagai alat bantu dalam klasifikasi citra mammografi [7]. Sejumlah penelitian menunjukkan bahwa CNN mampu memberikan performa yang baik dalam membedakan citra mammografi dengan kategori *benign* dan *malignant*, sehingga berpotensi digunakan sebagai alat bantu dalam proses klasifikasi kanker payudara[8], [9].

Sejumlah penelitian mengenai klasifikasi citra mammografi memanfaatkan dataset publik, seperti CBIS-DDSM, sebagai data pelatihan maupun pengujian model CNN [10], [11], [12]. Dataset publik memberikan kemudahan dalam pengembangan dan evaluasi model karena menyediakan data yang telah dianotasi serta banyak digunakan sebagai acuan dalam penelitian klasifikasi citra mammografi [10], [11], [12]. Namun, performa model yang diperoleh dari dataset publik belum tentu merepresentasikan kemampuan sistem ketika diterapkan pada data klinis yang berasal dari institusi kesehatan yang berbeda [13], [14]. Perbedaan karakteristik citra, protokol akuisisi, perangkat mammografi, dan populasi pasien dapat menyebabkan pergeseran distribusi data (*domain shift*) yang berpotensi memengaruhi kemampuan generalisasi model [13], [14], [15]. Permasalahan tersebut menunjukkan bahwa keberhasilan model dalam mengklasifikasikan citra pada dataset pengembangan belum cukup untuk menggambarkan konsistensi performanya pada sumber data lain. Oleh karena itu, evaluasi eksternal diperlukan untuk mengetahui sejauh mana model mampu menghasilkan prediksi yang sesuai ketika dihadapkan pada data yang tidak terlibat dalam proses pengembangan. Selain mengukur performa klasifikasi, evaluasi eksternal juga dapat digunakan untuk mengidentifikasi kesalahan prediksi pada masing-masing kelas sehingga memberikan gambaran mengenai keterbatasan model ketika diterapkan pada karakteristik data yang berbeda [13], [14].

Berdasarkan permasalahan tersebut, evaluasi menggunakan data dari rumah sakit mitra diperlukan untuk mengetahui performa model pada sumber data klinis yang berbeda dari lingkungan pengembangannya. Penelitian ini menggunakan 15 citra mammografi klinis yang telah dianonimkan dan memiliki label diagnosis, yang diperoleh dari RSUD Siti Fatimah Palembang sebagai rumah sakit mitra. Data tersebut digunakan sebagai data eksternal untuk menguji kesesuaian hasil prediksi model terhadap label diagnosis serta mengidentifikasi kesalahan klasifikasi. Penelitian ini bertujuan mengevaluasi performa model CNN EfficientNetB2 yang telah dikembangkan sebelumnya melalui penerapan pada sistem klasifikasi citra mammografi berbasis website lokal. Berbeda dengan penelitian yang berfokus pada pengembangan dan pelatihan model, penelitian ini menitikberatkan pada evaluasi eksternal terhadap model yang telah tersedia menggunakan data klinis dari sumber yang berbeda. Kontribusi penelitian ini terletak pada pengujian model menggunakan citra mammografi klinis dari rumah sakit mitra, analisis kesesuaian hasil prediksi terhadap label diagnosis, identifikasi kesalahan klasifikasi, serta penerapan model melalui sistem berbasis website lokal. Evaluasi dilakukan menggunakan *confusion matrix* dengan metrik *accuracy*, *precision*, *recall*, dan *F1-score* untuk memberikan gambaran awal mengenai performa model pada data mammografi klinis eksternal.

## 2. TINJAUAN PUSTAKA

*Convolutional Neural Network* (CNN) merupakan salah satu metode *Deep Learning* yang digunakan dalam pengolahan dan klasifikasi citra karena mampu mempelajari karakteristik visual dari data melalui proses ekstraksi fitur secara bertahap [16], [17]. CNN memiliki beberapa lapisan utama, yaitu *convolutional layer, pooling layer*, dan *fully connected layer* yang bekerja untuk mengekstraksi serta mengolah fitur citra hingga menghasilkan keluaran klasifikasi. Pada bidang pencitraan medis, CNN telah digunakan dalam klasifikasi citra mammografi dan menunjukkan kemampuan dalam mengenali pola visual yang berkaitan dengan karakteristik jaringan dan lesi payudara [16], [18]. Salah satu arsitektur CNN yang telah diterapkan dalam klasifikasi citra mammografi adalah EfficientNetB2. Penelitian sebelumnya menggunakan arsitektur tersebut pada dataset CBIS-DDSM untuk mengklasifikasikan citra mammografi berdasarkan karakteristik kanker payudara [9]. Selain penerapan CNN pada bidang pencitraan medis, penelitian lainnya juga membahas implementasi CNN pada sistem pengenalan wajah berbasis lokal, meskipun memiliki objek dan tujuan klasifikasi yang berbeda, penelitian tersebut dapat menjadi referensi pendukung terkait penerapan model CNN ke dalam suatu sistem berbasis lokal [19]. Dalam penelitian ini, CNN digunakan untuk mengklasifikasikan citra mammografi ke dalam dua kategori, yaitu *Benign* dan *Malignant*, berdasarkan karakteristik visual yang terdapat pada citra.

EfficientNet merupakan salah satu arsitektur CNN yang dikembangkan dengan mempertimbangkan keseimbangan antara akurasi dan efisiensi komputasi [20]. Arsitektur ini menggunakan pendekatan compound scaling yang melakukan penskalaan secara seimbang terhadap kedalaman jaringan, lebar jaringan, dan resolusi citra [20]. EfficientNet memiliki beberapa varian, salah satunya EfficientNetB2, yang menerapkan prinsip compound scaling dengan konfigurasi kedalaman, lebar jaringan, dan resolusi input yang disesuaikan [20]. Dalam penelitian ini, arsitektur EfficientNetB2 digunakan sebagai model klasifikasi yang menerima citra mammografi berukuran 320 × 320 piksel untuk menghasilkan prediksi ke dalam kelas Benign dan Malignant. Pemilihan arsitektur tersebut didasarkan pada penelitian sebelumnya yang menerapkan EfficientNetB2 untuk klasifikasi citra mammografi menggunakan dataset CBIS-DDSM [9]. Berbeda dengan penelitian sebelumnya yang berfokus pada pengembangan dan penerapan model *Fine-Tuned* EfficientNetB2, penelitian ini menggunakan model yang telah dikembangkan sebelumnya tanpa melakukan pelatihan ulang (*retraining*) maupun penyesuaian model (*fine-tuning*). Oleh karena itu, fokus penelitian diarahkan pada evaluasi eksternal terhadap model menggunakan citra mammografi klinis yang berasal dari sumber data berbeda melalui sistem klasifikasi berbasis website lokal.

Evaluasi eksternal merupakan pengujian model menggunakan data yang berasal dari sumber berbeda dan tidak dilibatkan dalam proses pengembangan model [13], [14]. Perbedaan karakteristik data dan lingkungan klinis dapat memengaruhi performa serta kemampuan generalisasi model ketika diterapkan pada sumber data lain [14], [15]. Kondisi tersebut dapat terjadi akibat pergeseran distribusi data (*domain shift*), misalnya karena perbedaan karakteristik citra, protokol akuisisi, perangkat mammografi, dan populasi pasien [14], [15]. Dalam penelitian ini, evaluasi eksternal dilakukan menggunakan 15 citra mammografi klinis yang telah dianonimkan dari RSUD Siti Fatimah Palembang. Hasil prediksi model melalui sistem klasifikasi berbasis website lokal dibandingkan dengan label diagnosis pada data asli untuk mengetahui kesesuaian prediksi dan mengidentifikasi kesalahan klasifikasi. Performa model klasifikasi dapat dievaluasi menggunakan *accuracy, precision, recall* dan F1-*score*. *Accuracy* menunjukkan proporsi prediksi yang benar terhadap seluruh data, *precision* menunjukkan ketepatan model dalam menghasilkan prediksi positif, sedangkan *recall* menunjukkan kemampuan model dalam mengidentifikasi data positif dan F1-*score* merupakan rata-rata harmonis antara *precision* dan *recall* [21]. Perhitungan *accuracy, precision, recall*, dan F1-*score* dapat dinyatakan menggunakan persamaan berikut [21]:

$$Accuracy = \frac{TP + TN}{TP+TN+FP+FN} \quad \text{.....(2.1)}$$

$$Precision = \frac{TP}{TP+FP} \quad \text{.....(2.2)}$$

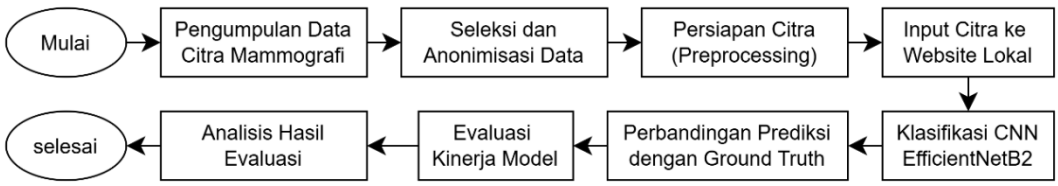
$$Recall = \frac{TP}{TP+FN} \quad \text{.....(2.3)}$$

$$\text{F1-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \text{.....(2.4)}$$

Dengan TP (*True Positive*), TN (*True Negative*), FP (*False Positive*), dan FN (*False Negative*) merupakan komponen *confusion matrix*. TP menunjukkan jumlah data positif yang diprediksi positif, TN menunjukkan jumlah data negatif yang diprediksi negatif, FP menunjukkan jumlah data negatif yang diprediksi positif, sedangkan FN menunjukkan jumlah data positif yang diprediksi negatif. Dalam penelitian ini, kelas *Malignant* ditetapkan sebagai kelas positif, sedangkan *Benign* sebagai kelas negatif. Keempat komponen tersebut digunakan untuk menghitung metrik evaluasi berdasarkan perbandingan antara hasil prediksi model dan label diagnosis sebenarnya. Penetapan kelas positif dan negatif tersebut menjadi dasar dalam interpretasi hasil evaluasi. Dengan demikian, nilai *precision* menunjukkan ketepatan prediksi *Malignant* terhadap seluruh citra yang diprediksi *Malignant*, sedangkan *recall* menunjukkan proporsi citra *Malignant* yang berhasil dikenali dari seluruh citra dengan label *Malignant* sebenarnya. Kesalahan klasifikasi juga dapat diidentifikasi melalui nilai *False Positive* dan *False Negative* untuk mengetahui jenis ketidaksesuaian prediksi yang dihasilkan model.

## 3. METODE PENELITIAN

Penelitian ini menggunakan metode eksperimen dengan pendekatan kuantitatif untuk menganalisis performa model Convolutional Neural Network (CNN) melalui evaluasi eksternal pada sistem klasifikasi citra mammografi berbasis website lokal. Model EfficientNetB2 yang digunakan dalam penelitian ini merupakan model yang telah dikembangkan pada penelitian tugas akhir sebelumnya menggunakan pendekatan transfer learning pada dataset CBIS-DDSM. Pada penelitian ini, model tersebut digunakan kembali tanpa dilakukan pelatihan ulang maupun perubahan arsitektur untuk mengevaluasi kemampuannya ketika diterapkan pada data eksternal. Dengan demikian, penelitian ini berfokus pada pengujian hasil klasifikasi model terhadap data klinis yang berasal dari sumber berbeda, bukan pada pengembangan atau optimasi ulang model. Tahapan penelitian meliputi pengumpulan dan seleksi data mammografi klinis, persiapan data, penerapan model pada sistem berbasis website lokal, pengujian klasifikasi setiap citra, perbandingan hasil prediksi dengan label diagnosis asli, serta analisis hasil evaluasi. Tahapan tersebut disajikan pada Gambar 1.




Gambar 1. Alur Penelitian

Gambar 1 menunjukkan proses evaluasi eksternal yang diawali dengan pengumpulan dan seleksi citra mammografi klinis. Citra terpilih kemudian dimasukkan ke dalam sistem klasifikasi berbasis website lokal untuk memperoleh prediksi kelas *Benign* atau *Malignant* menggunakan model EfficientNetB2. Hasil prediksi selanjutnya dibandingkan dengan label diagnosis asli (*ground truth*) untuk mengetahui

kesesuaian klasifikasi. Tahapan akhir adalah analisis hasil evaluasi untuk mengetahui performa model dalam mengklasifikasikan citra mammografi klinis eksternal.

Data eksternal berupa 15 citra mammografi klinis yang diperoleh dari bagian Radiologi RSUD Siti Fatimah Palembang. Citra tersebut berasal dari arsip pemeriksaan mammografi sejak Januari 2026 dan masing-masing mewakili satu pasien yang berbeda. Data diperoleh berdasarkan izin penggunaan untuk keperluan penelitian dan seluruh citra telah dianonimkan sebelum digunakan. Pemilihan citra dilakukan berdasarkan penilaian dan rekomendasi tenaga medis terhadap kondisi serta kualitas citra. Citra dengan tingkat *noise* yang tinggi atau objek/lesi berukuran terlalu kecil sehingga detailnya sulit diamati tidak dipilih untuk pengujian. Berdasarkan proses seleksi tersebut, diperoleh 15 citra yang terdiri atas 7 citra berlabel *Benign* dan 8 citra berlabel *Malignant*. Label tersebut mengacu pada hasil diagnosis radiolog yang tercantum dalam arsip pemeriksaan asli dan digunakan sebagai *ground truth* dalam evaluasi. Seluruh citra merupakan data eksternal yang tidak digunakan dalam proses pengembangan maupun pelatihan model EfficientNetB2 sebelumnya.

Pengujian dilakukan dengan memasukkan setiap citra mammografi secara bergantian ke dalam website lokal yang telah terintegrasi dengan model EfficientNetB2 hasil pengembangan penelitian sebelumnya. Model tersebut dikembangkan menggunakan pendekatan *transfer learning* pada dataset CBIS-DDSM dengan bobot awal *pre-trained ImageNet*. Model digunakan kembali tanpa pelatihan ulang maupun perubahan arsitektur. Website berfungsi sebagai antarmuka pengujian yang menerima masukan citra, meneruskannya ke model untuk proses inferensi, dan menampilkan hasil klasifikasi. Sebelum diklasifikasikan, citra diproses melalui tahapan *preprocessing* sesuai dengan *pipeline* yang digunakan pada model sebelumnya, meliputi normalisasi intensitas piksel dan penyesuaian ukuran citra menjadi 320 × 320 piksel. Selanjutnya, model menghasilkan prediksi kelas *Benign* atau *Malignant* berdasarkan citra yang telah melalui tahapan *preprocessing*. Dengan demikian, website digunakan sebagai media penerapan dan pengujian model, bukan sebagai objek utama pengembangan dalam penelitian ini. Hasil prediksi setiap citra dibandingkan dengan *ground truth* yang berasal dari diagnosis radiolog pada arsip pemeriksaan asli. Prediksi dinyatakan sesuai apabila kelas hasil klasifikasi sama dengan label diagnosis, sedangkan prediksi dinyatakan tidak sesuai apabila keduanya berbeda. Evaluasi kuantitatif dilakukan menggunakan *confusion matrix* untuk memperoleh nilai *accuracy*, *precision*, *recall*, dan F1-*score*. Selain itu, hasil pengujian, khususnya citra dengan prediksi yang tidak sesuai dengan diagnosis asli, didiskusikan bersama tenaga medis untuk meninjau karakteristik citra dan memperoleh penjelasan medis mengenai kemungkinan penyebab ketidaksesuaian tersebut. Masukan dari diskusi digunakan sebagai bahan pendukung interpretasi hasil evaluasi dan tidak mengubah label diagnosis asli yang menjadi *ground truth*.

*Tabel 1. Komponen Evaluasi Penelitian*

| Komponen | Keterangan |
|---|---|
| Sistem | Website lokal klasifikasi citra mammografi |
| Model CNN | EfficientNetB2 (hasil pengembangan penelitian sebelumnya) |
| Data Evaluasi | 15 citra mammografi klinis anonim dari 15 pasien berbeda |
| Sumber Data | Bagian Radiologi RSUD Siti Fatimah |
| Jenis Evaluasi | Evaluasi eksternal |
| *Output* Sistem | *Benign* dan *Malignant* |
| *Ground Truth* | Label diagnosis berdasarkan diagnosis dokter pada data klinis asli |
| Bentuk Evaluasi | Perbandingan hasil prediksi dengan *ground truth* |

Tabel 1 merangkum komponen utama yang digunakan dalam evaluasi eksternal, meliputi sistem, model, data, dan bentuk evaluasi yang dilakukan. Tingkat kesesuaian prediksi dihitung berdasarkan perbandingan jumlah prediksi yang sesuai terhadap seluruh citra eksternal menggunakan Persamaan (3.1).

$$\text{Tingkat Kesesuaian} = \frac{\text{Jumlah Prediksi Sesuai}}{\text{Jumlah Seluruh Data Eksternal}} \times 100\% \quad \text{(3.1)}$$

DOI: 10.36257/teliska.v19i3.12682
Jurnal Teliska, Vol.19 No. 3, Tahun 2026

Hasil perhitungan tingkat kesesuaian, confusion matrix, dan metrik evaluasi lainnya disajikan pada bagian Hasil dan Pembahasan.

## 4. HASIL DAN PEMBAHASAN

Penelitian ini menganalisis performa model *Convolutional Neural Network* (CNN) berbasis EfficientNetB2 melalui evaluasi eksternal pada sistem klasifikasi citra mammografi berbasis website lokal. Model EfficientNetB2 yang digunakan merupakan model yang telah dikembangkan sebelumnya, sehingga penelitian ini tidak melakukan pelatihan ulang maupun perubahan arsitektur. Analisis difokuskan pada hasil klasifikasi model ketika diterapkan pada data di luar lingkungan pengembangannya, sedangkan website lokal digunakan sebagai media penerapan dan pengujian model. Pada tahap pengembangan, model EfficientNetB2 menghasilkan *accuracy* sebesar 72%, *precision* sebesar 67%, *recall* sebesar 96% dan F1-*score* sebesar 79%. Nilai tersebut digunakan sebagai konteks terhadap model yang dievaluasi pada penelitian ini. Hasil pada tahap pengembangan tidak dibandingkan secara langsung dengan tingkat kesesuaian prediksi pada data eksternal karena keduanya diperoleh dari data dan prosedur evaluasi yang berbeda. Penyajian hasil tersebut bertujuan memberikan gambaran mengenai performa model sebelum diterapkan pada data eksternal. Evaluasi eksternal dilakukan menggunakan 15 citra mammografi klinis anonim yang tidak terlibat dalam proses pengembangan model. Setiap citra diunggah melalui website lokal untuk memperoleh hasil klasifikasi berupa kelas *Benign* atau *Malignant*. Hasil prediksi kemudian dibandingkan dengan label diagnosis pada data asli sebagai *ground truth*. Prediksi dinyatakan sesuai apabila kelas hasil sistem sama dengan *ground truth*, sedangkan prediksi dinyatakan tidak sesuai apabila keduanya berbeda. Hasil pengujian masing-masing citra ditunjukkan pada Tabel 2.

*Tabel 2. Hasil Pengujian pada Data Eksternal*

| No | ID Citra | *Ground Truth* | Prediksi Sistem | Keterangan |
|---|---|---|---|---|
| 1 | Citra-01 | *Benign* | *Benign* | Sesuai |
| 2 | Citra-02 | *Benign* | *Benign* | Sesuai |
| 3 | Citra-03 | *Benign* | *Malignant* | Tidak Sesuai |
| 4 | Citra-04 | *Benign* | *Malignant* | Tidak Sesuai |
| 5 | Citra-05 | *Benign* | *Benign* | Sesuai |
| 6 | Citra-06 | *Benign* | *Benign* | Sesuai |
| 7 | Citra-07 | *Benign* | *Benign* | Sesuai |
| 8 | Citra-08 | *Malignant* | *Malignant* | Sesuai |
| 9 | Citra-09 | *Malignant* | *Malignant* | Sesuai |
| 10 | Citra-10 | *Malignant* | *Malignant* | Sesuai |
| 11 | Citra-11 | *Malignant* | *Malignant* | Sesuai |
| 12 | Citra-12 | *Malignant* | *Malignant* | Sesuai |
| 13 | Citra-13 | *Malignant* | *Malignant* | Sesuai |
| 14 | Citra-14 | *Malignant* | *Malignant* | Sesuai |
| 15 | Citra-15 | *Malignant* | *Malignant* | Sesuai |

Berdasarkan Tabel 2, dari 15 citra mammografi eksternal yang diuji, sebanyak 13 citra menghasilkan prediksi yang sesuai dengan *ground truth*, sedangkan 2 citra menghasilkan prediksi yang tidak sesuai. Tingkat kesesuaian prediksi dihitung berdasarkan perbandingan jumlah prediksi yang sesuai terhadap seluruh data eksternal menggunakan Persamaan (3.1).

$$\text{Tingkat Kesesuaian} = \frac{\text{Jumlah Prediksi Sesuai}}{\text{Jumlah Seluruh Data Eksternal}} \times 100\%$$

$$\text{Tingkat Kesesuaian} = \frac{13}{15} \times 100 = 86{,}67\%$$

Berdasarkan perhitungan menggunakan Persamaan (3.1), diperoleh tingkat kesesuaian prediksi pada data eksternal sebesar 86,67%. Confusion matrix untuk data eksternal disajikan pada Tabel 3.

*Tabel 3. Confusion Matrix Hasil Evaluasi Eksternal*

| *Ground Truth* | Prediksi *Benign* | Prediksi *Malignant* |
|---|---|---|
| *Benign* | 5 (TN) | 2 (FP) |
| *Malignant* | 0 (FN) | 8 (TP) |

Berdasarkan Tabel 3, dengan kelas *Malignant* sebagai kelas positif, diperoleh 8 *True Positive* (TP), 5 *True Negative* (TN), 2 *False Positive* (FP), dan 0 *False Negative* (FN). Komponen tersebut selanjutnya digunakan untuk menghitung *accuracy*, *precision*, *recall*, dan F1-*score* berdasarkan Persamaan (2.1)–(2.4).

$$\begin{aligned} Accuracy &= \frac{TP + TN}{TP+TN+FP+FN} \\ &= \frac{8 + 5}{8+5+2+0} \\ &= 0{,}8667 \times 100\% \\ &= 86{,}67\% \end{aligned}$$

$$\begin{aligned} Precision &= \frac{TP}{TP+FP} \\ &= \frac{8}{8+2} \\ &= 0{,}80 \times 100\% \\ &= 80\% \end{aligned}$$

$$\begin{aligned} Recall &= \frac{TP}{TP+FN} \\ &= \frac{8}{8+0} \\ &= 1{,}00 \times 100\% \\ &= 100\% \end{aligned}$$

$$\begin{aligned} \text{F1-}Score &= 2 \times \frac{Precision \times Recall}{Precision + Recall} \\ &= 2 \times \frac{0{,}80 \times 1{,}00}{0{,}80 + 1{,}00} \\ &= 0{,}8889 \times 100\% \\ &= 88{,}89\% \end{aligned}$$

Hasil perhitungan menunjukkan bahwa evaluasi eksternal menghasilkan *accuracy* sebesar 86,67%, *precision* sebesar 80%, *recall* sebesar 100%, dan F1-*score* sebesar 88,89%. Hasil tersebut menunjukkan bahwa seluruh citra dengan label *Malignant* pada data eksternal berhasil diklasifikasikan sebagai *Malignant*, sedangkan dua citra dengan label *Benign* diklasifikasikan sebagai *Malignant*. Dengan demikian, kesalahan klasifikasi pada data eksternal seluruhnya berupa *False Positive*, sedangkan tidak terdapat *False Negative* pada data yang diuji.

*Tabel 4. Perbandingan Hasil Evaluasi Model pada Lingkungan Pengembangan dan Website Lokal*

| Metrik | Lingkungan Pengembangan | Website Lokal pada Data Eksternal |
|---|---|---|
| *Accuracy* | 72% | 86,67% |
| *Precision* | 67% | 80% |
| *Recall* | 96% | 100% |
| *F1-score* | 79% | 88,89% |

Berdasarkan Tabel 4, terdapat perbedaan hasil evaluasi model EfficientNetB2 antara lingkungan pengembangan dan penerapan pada website lokal menggunakan data eksternal. Pada lingkungan

pengembangan, model memperoleh *accuracy* sebesar 72%, *precision* sebesar 67%, *recall* sebesar 96%, dan F1-*score* sebesar 79%. Sementara itu, pada penerapan melalui website lokal menggunakan data eksternal, diperoleh *accuracy* sebesar 86,67%, *precision* sebesar 80%, *recall* sebesar 100%, dan F1-*score* sebesar 88,89%. Perbedaan tersebut tidak digunakan untuk menyatakan peningkatan atau penurunan performa model secara langsung karena kedua hasil diperoleh menggunakan data dan konteks evaluasi yang berbeda. Perbandingan ini digunakan untuk menunjukkan hasil penerapan model pada lingkungan website lokal ketika diuji menggunakan data yang berada di luar lingkungan pengembangannya.

Perbedaan hasil evaluasi tersebut perlu diinterpretasikan dengan mempertimbangkan sumber data, jumlah data, dan kondisi pengujian pada masing-masing penelitian. Pada tahap pengembangan model, evaluasi dilakukan menggunakan data pengujian dari dataset CBIS-DDSM yang berasal dari *The Cancer Imaging Archive* (TCIA), dengan jumlah 332 citra dari total 2.201 citra mammografi yang berasal dari 977 pasien unik. Sementara itu, penelitian ini menggunakan 15 citra mammografi klinis dari RSUD Siti Fatimah Palembang yang berasal dari 15 pasien berbeda. Model EfficientNetB2 yang telah dikembangkan sebelumnya diterapkan pada data eksternal tersebut tanpa pelatihan ulang maupun perubahan arsitektur. Dengan demikian, kedua hasil evaluasi diperoleh dari sumber data, jumlah sampel, dan kondisi pengujian yang berbeda. Perbedaan karakteristik serta distribusi data tersebut dapat memengaruhi hasil klasifikasi. Oleh karena itu, perbandingan metrik pada Tabel 4 hanya digunakan untuk memberikan gambaran performa model pada lingkungan pengembangan dan pengujian eksternal, bukan untuk menyimpulkan adanya peningkatan atau penurunan performa secara langsung.

Dua prediksi yang tidak sesuai pada Tabel 2 terjadi pada Citra-03 dan Citra-04, yang memiliki *ground truth Benign* tetapi diprediksi sebagai *Malignant*. Kondisi tersebut merupakan *False Positive* dan menunjukkan bahwa model masih dapat menghasilkan kesalahan klasifikasi pada data eksternal. Secara umum, perbedaan karakteristik citra, kualitas citra, perangkat atau prosedur akuisisi, serta distribusi data dapat menjadi faktor yang memengaruhi keluaran model ketika diterapkan pada lingkungan yang berbeda. Namun, penelitian ini tidak melakukan analisis khusus terhadap karakteristik kedua citra tersebut, sehingga penyebab spesifik kesalahan klasifikasi tidak dapat ditentukan berdasarkan hasil yang diperoleh. Kesalahan klasifikasi berupa False Positive berpotensi menyebabkan citra dengan label Benign teridentifikasi sebagai Malignant. Dalam konteks pemeriksaan medis, kondisi tersebut dapat menimbulkan kebutuhan pemeriksaan lanjutan atau konfirmasi diagnosis oleh tenaga medis. Sebaliknya, tidak ditemukannya False Negative pada 15 citra yang diuji menunjukkan bahwa seluruh citra dengan label Malignant berhasil dikenali oleh model dalam pengujian ini. Meskipun demikian, hasil tersebut tidak dapat dijadikan jaminan bahwa model tidak akan menghasilkan False Negative pada data yang lebih luas. Oleh karena itu, hasil klasifikasi model hanya dapat digunakan sebagai informasi pendukung dan tidak menggantikan interpretasi serta diagnosis radiolog. Hasil evaluasi eksternal memberikan gambaran mengenai kemampuan generalisasi model ketika diterapkan pada data di luar lingkungan pengembangannya. Literatur mengenai evaluasi eksternal pada kecerdasan buatan untuk pencitraan medis menekankan bahwa pengujian menggunakan data dari lingkungan berbeda diperlukan untuk menilai kemampuan model dalam mempertahankan performanya pada kondisi yang tidak digunakan selama pengembangan [13], [14]. Perbedaan karakteristik citra, protokol akuisisi, perangkat mammografi, dan populasi dapat menyebabkan pergeseran distribusi data atau *domain shift* yang berpotensi memengaruhi hasil klasifikasi [13], [14], [15]. Dalam penelitian ini, sebagian besar citra eksternal berhasil diklasifikasikan sesuai dengan label diagnosis aslinya. Namun, ditemukannya dua False Positive menunjukkan bahwa model masih memiliki keterbatasan dalam mengklasifikasikan sebagian citra dari sumber data eksternal.

Dari sisi penerapan, website lokal berfungsi sebagai media untuk menjalankan model EfficientNetB2 pada data eksternal secara langsung dan terstruktur. Sistem menerima citra mammografi, menghasilkan klasifikasi *Benign* atau *Malignant*, kemudian hasil tersebut dibandingkan dengan *ground truth*. Dengan demikian, website lokal mendukung proses implementasi dan pengujian model, sedangkan fokus utama penelitian tetap berada pada evaluasi hasil klasifikasi terhadap data eksternal. Meskipun hasil evaluasi

menunjukkan tingkat kesesuaian sebesar 86,67%, hasil tersebut belum dapat digunakan untuk menyatakan bahwa model memiliki performa klinis yang tinggi secara umum. Jumlah data eksternal yang hanya terdiri atas 15 citra masih terbatas dan belum cukup untuk merepresentasikan variasi data mammografi secara luas. Selain itu, data eksternal yang digunakan dalam penelitian ini belum mencakup variasi antar-institusi, perangkat, prosedur akuisisi, dan populasi secara menyeluruh. Oleh karena itu, hasil penelitian ini perlu dipandang sebagai gambaran awal mengenai kemampuan generalisasi model pada data eksternal yang diuji. Secara keseluruhan, evaluasi eksternal menunjukkan bahwa model EfficientNetB2 dapat diterapkan pada sistem klasifikasi citra mammografi berbasis website lokal dan menghasilkan klasifikasi yang sesuai pada sebagian besar data eksternal yang diuji. Hasil evaluasi juga menunjukkan adanya dua *False Positive*, sehingga kemampuan generalisasi model masih memiliki keterbatasan pada sebagian citra. Temuan tersebut menunjukkan pentingnya penggunaan data eksternal dalam mengevaluasi kemampuan model di luar lingkungan pengembangannya.

## 5. KESIMPULAN

Penelitian ini menunjukkan bahwa model *Convolutional Neural Network* (CNN) berbasis EfficientNetB2 yang telah dikembangkan sebelumnya dapat diterapkan pada sistem klasifikasi citra mammografi berbasis website lokal dan dievaluasi menggunakan data eksternal. Evaluasi terhadap 15 citra mammografi klinis anonim menunjukkan bahwa 13 citra menghasilkan prediksi yang sesuai dengan label diagnosis, sedangkan 2 citra menghasilkan prediksi yang tidak sesuai, dengan tingkat kesesuaian prediksi sebesar 86,67%. Berdasarkan hasil *confusion matrix*, diperoleh *accuracy* sebesar 86,67%, *precision* sebesar 80%, *recall* sebesar 100%, dan F1-*score* sebesar 88,89%. Hasil tersebut menunjukkan bahwa model mampu menghasilkan klasifikasi yang sesuai pada sebagian besar data eksternal yang diuji, namun masih terdapat keterbatasan generalisasi yang ditunjukkan oleh dua *False Positive*. Kelebihan penelitian ini terletak pada penerapan model pada data yang tidak terlibat dalam proses pengembangan melalui sistem website lokal, sehingga memberikan gambaran awal mengenai kemampuan model ketika diterapkan pada data di luar lingkungan pengembangannya. Keterbatasan penelitian terdapat pada jumlah data eksternal yang masih terbatas, yaitu 15 citra, sehingga belum dapat merepresentasikan variasi data mammografi secara luas. Pengembangan selanjutnya dapat dilakukan dengan menggunakan data eksternal yang lebih banyak dan lebih beragam serta melibatkan data dari beberapa institusi untuk memperoleh gambaran kemampuan generalisasi model yang lebih kuat.

## DAFTAR PUSTAKA

[1] J. S. Brown, S. R. Amend, R. H. Austin, R. A. Gatenby, E. U. Hammarlund, and K. J. Pienta, "Updating the Definition of Cancer.," *Mol. Cancer Res.*, vol. 21, no. 11, pp. 1142–1147, Nov. 2023, doi: 10.1158/1541-7786.MCR-23-0411.

[2] F. Bray *et al.*, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA. Cancer J. Clin.*, vol. 74, no. 3, pp. 229–263, 2024, doi: 10.3322/caac.21834.

[3] A. M. Filho *et al.*, "The GLOBOCAN 2022 cancer estimates: Data sources, methods, and a snapshot of the cancer burden worldwide," *Int. J. Cancer*, vol. 156, no. 7, pp. 1336–1346, 2025, doi: 10.1002/ijc.35278.

[4] W. H. Organization, "Breast cancer," World Health Organization. Accessed: Aug. 05, 2026. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[5] K. Loizidou, R. Elia, and C. Pitris, "Computer-aided breast cancer detection and classification in mammography: A comprehensive review," *Comput. Biol. Med.*, vol. 153, no. December 2022, p. 106554, 2023, doi: 10.1016/j.compbiomed.2023.106554.

[6] N. Chotai, A. Gadwal, D. Buchireddy, and W. T. Yang, "Why we still miss breast cancers : strategies for improving mammography interpretation," *Insights Imaging*, vol. 17, no. 1, p. 8, 2026, doi: 10.1186/s13244-025-02168-2.
[7] L. Wang, "Mammography with deep learning for breast cancer detection," *Front. Oncol.*, vol. 14, no. February, pp. 1–16, 2024, doi: 10.3389/fonc.2024.1281922.
[8] Z. A. Ansari, M. M. Tripathi, and R. Ahmed, "The role of explainable AI in enhancing breast cancer diagnosis using machine learning and deep learning models," *Discov. Artif. Intell.*, vol. 5, no. 1, 2025, doi: 10.1007/s44163-025-00307-8.
[9] V. K. V., A. N. Nayak, and T. R. Arunkumar, "Sensitivity-Based Breast Cancer Characterization on CBIS-DDSM Mammography Dataset Using Fine-Tuned EfficientNet-B2 Model," *Int. J. Adv. Electr. Comput. Eng.*, vol. 15, no. 1S, pp. 365–371, 2026, doi: 10.65521/ijaece.v15i1s.1380.
[10] J. Logan, P. J. Kennedy, and D. Catchpoole, "A review of the machine learning datasets in mammography, their adherence to the FAIR principles and the outlook for the future," *Sci. Data*, vol. 10, no. 1, pp. 1–11, 2023, doi: 10.1038/s41597-023-02430-6.
[11] A. Mračko, L. Vanovčanová, and I. Cimrák, "Mammography Datasets for Neural Networks—Survey," *J. Imaging*, vol. 9, no. 5, 2023, doi: 10.3390/jimaging9050095.
[12] H. R. Fajrin and S. D. Min, "From Machine Learning to Ensemble Approaches: A Systematic Review of Mammogram Classification Methods," *Diagnostics*, vol. 15, no. 22, pp. 1–30, 2025, doi: 10.3390/diagnostics15222829.
[13] A. C. Yu, B. Mohajer, and J. Eng, "External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review," *Radiol. Artif. Intell.*, vol. 4, no. 3, pp. 1–9, 2022, doi: 10.1148/ryai.210064.
[14] M. U. Suleman *et al.*, "Assessing the generalizability of artificial intelligence in radiology: a systematic review of performance across different clinical settings," *Ann. Med. Surg.*, vol. 87, no. 12, pp. 8803–8811, 2025, doi: 10.1097/ms9.0000000000004166.
[15] S. Matta *et al.*, "A systematic review of generalization research in medical image classification," *Comput. Biol. Med.*, vol. 183, 2024, doi: 10.1016/j.compbiomed.2024.109256.
[16] S. Chakravarthy *et al.*, "Multi-class Breast Cancer Classification Using CNN Features Hybridization," *Int. J. Comput. Intell. Syst.*, vol. 17, no. 1, 2024, doi: 10.1007/s44196-024-00593-7.
[17] D. Moslemi, S. M. Hassan Hosseini, E. Jafarian, and M. Jamshidi, "Benchmarking Deep Learning Algorithms for Breast Cancer Detection: A Comprehensive Review and Evaluation Across Public Imaging Datasets," *Infosci. Trends*, vol. 2, no. 4, pp. 11–24, 2025, doi: 10.61186/ist.202502.04.02.
[18] G. W. Purnama, A. Aan, J. Permana, and N. Ketut, "Klasifikasi Citra Medis Mamografi Berbasis Multi-View Convolutional Neural Networks," *J. Pendidik. Teknol. dan Kejuru.*, vol. 22, no. 2, pp. 162–171, 2025, doi: 10.23887/jptk-undiksha.v22i2.94691.
[19] D. P. Sari, M. Arya, and C. Putra, "Implementasi Pengenalan Wajah Berbasis CNN dan RFID untuk Area Akses Aman di Fasilitas Polsri dengan Hosting Lokal," *J. Teliska*, vol. 17, no. II, pp. 23–31, 2024, doi: 10.5281/zenodo.13047545.
[20] M. Tan and Q. V Le, "EfficientNet : Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research (PMLR), 2019, pp. 6105–6114. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html
[21] N. I. Fardana, R. R. Isnanto, and O. D. Nurhayati, "Pneumothorax Detection System in Thoracic Radiography Images Using CNN Method," *Sci. J. Informatics*, vol. 11, no. 4, pp. 981–990, 2025, doi: 10.15294/sji.v11i4.16635.