Indonesian Journal of Digital Business

Journal homepage: https://ejournal.upi.edu/index.php/IJDB

Pengembangan Sistem Dermatologi Veteriner Kucing dan Anjing Berbasis Two-Stage Lesion-Aware Pipeline dengan Mekanisme Confidence Calibration

Assariy Ramdhani[1], Randi Proska Sandra[2] , Vikri Aulia[3], Dony Novaliendry[4]

[1]Program Studi Informatika, Fakultas Teknik, Universitas Negeri Padang, Indonesia
[2,3,4]Universitas Negeri Padang, Indonesia
Correspondence: E-mail: assariyramdhani@gmail.com, randiproska@ft.unp.ac.id, vikriaulia@unp.ac.id, Dony.Novaliendry@ft.unp.ac.id

A B S T R A C T

Tingginya prevalensi penyakit kulit anjing dan kucing di Indonesia serta kemiripan visual antarkondisi sering memicu misdiagnosis. Mengatasi kelemahan detektor konvensional pada lesi kecil, penelitian ini mengembangkan *Two-Stage Lesion-Aware Pipeline* berbasis YOLOv12 untuk lokalisasi dan EfficientNetV2-S untuk klasifikasi sebagai alat penapisan dermatologi hewan terintegrasi AI. Sistem ini diperkuat dengan *Temperature Scaling* untuk kalibrasi kepercayaan model dan diimplementasikan pada aplikasi *mobile* berbasis Flutter/FastAPI. Hasil evaluasi menunjukkan performa mAP@0.5 sebesar 0,89 (kucing) dan 0,64 (anjing), serta *F1-score weighted* klasifikasi mencapai 0,9976 dan 0,9871. Secara signifikan, kalibrasi ($T = 1{,}3174$) berhasil mereduksi *Expected Calibration Error* (ECE) sebesar 32,4% menjadi 0,0209, memenuhi target fungsional ($ECE < 0{,}15$) dengan rata-rata waktu respons sistem 4,97 detik. Evaluasi usabilitas mencatat skor *Excellent* (85), sementara penilaian pakar memvalidasi kualitas antarmuka (4,44/5,00). *Pipeline* terkalibrasi ini terbukti layak sebagai instrumen penapisan lini pertama, dengan rekomendasi perluasan cakupan penyakit dan validasi pada *test set* independen sebelum diadopsi secara klinis.

A R T I C L E I N F O

Article History:
Submitted/Received 7 July 2026
First Revised 8 July 2026
Accepted 11 July 2026
First Available online 13 July 2026
Publication Date 13 July 2026

Keyword:
dermatologi veteriner;
Two-stage lesion-aware pipeline;
YOLOv12;
EfficientNetV2;
confidence calibration.

1. PENDAHULUAN

Indonesia, sebagai negara tropis dengan kelembapan tinggi, menyediakan ekosistem yang ideal bagi pertumbuhan patogen kulit, sehingga penyakit kulit menjadi salah satu alasan utama kunjungan ke klinik hewan. Data menunjukkan bahwa prevalensi dermatofitosis pada kucing di Indonesia mencapai 56,7% (Zaki et al., 2021), jauh lebih tinggi dibandingkan negara subtropis seperti Turki yang hanya 12,1% (Zaki et al., 2021). Tingginya insidensi ini bukan sekadar masalah kesejahteraan hewan, melainkan juga isu kesehatan masyarakat yang serius karena agen penyebab utamanya, *Microsporum canis*, dapat menular ke manusia. Situasi ini menciptakan urgensi nasional bagi teknologi deteksi dini yang cepat, akurat, dan mudah diakses oleh pemilik hewan guna memutus rantai penularan di tingkat masyarakat.

Diagnosis klinis dalam dermatologi veteriner menghadapi tantangan besar berupa mimikri visual, di mana penyakit dengan etiologi berbeda menunjukkan tanda klinis yang sangat mirip. Lesi berbentuk cincin yang identik dapat ditemukan baik pada infeksi jamur maupun bakteri, sehingga menyulitkan dokter hewan membedakannya hanya melalui inspeksi visual tanpa pemeriksaan laboratorium (Bard et al., 2023). Ketergantungan pada penilaian subjektif klinisi sering kali menyebabkan misdiagnosis dan penanganan yang tidak tepat. Sebuah sistem Computer-Aided Diagnosis (CAD) yang mampu menganalisis fitur morfologis lesi secara objektif dan konsisten sangat diperlukan untuk berfungsi sebagai pendapat kedua berbasis data yang mengurangi ambiguitas diagnostik dalam praktik klinis harian.

Adopsi telemedicine veteriner yang berkembang pesat di Indonesia belum diimbangi dengan peningkatan kualitas gambar yang sepadan dari pemilik hewan; foto yang dikirimkan sering kali buram, minim cahaya, atau diambil dari sudut yang kurang informatif, sehingga menurunkan tingkat kepercayaan dokter hewan terhadap diagnosis jarak jauh. Pendekatan deep learning standar berbasis *One-Stage Classification* juga memiliki kelemahan fundamental dalam mendeteksi lesi kulit yang kecil atau tersebar, karena seluruh gambar diproses sekaligus dan fitur lesi yang halus sering kali tertutupi oleh noise latar belakang yang kompleks. Untuk mengatasi hal ini, penelitian ini mengusulkan arsitektur *Two-Stage Lesion-Aware Pipeline*: tahap pertama melokalisasi dan memotong (crop) area lesi untuk menghilangkan noise latar belakang, dan tahap kedua melakukan analisis penyakit secara mendalam khusus pada area lesi yang telah dipotong tersebut. Pendekatan bertahap ini secara empiris terbukti meningkatkan sensitivitas deteksi secara signifikan, terutama untuk infeksi tahap awal yang sering terlewat (Abudukelimu et al., 2025).

Kendala lain terhadap adopsi klinis AI adalah sifatnya yang black-box serta kecenderungannya untuk *overconfidence* jaringan saraf sering mengeluarkan skor probabilitas yang sangat tinggi bahkan untuk prediksi yang salah, yang sangat berbahaya dalam konteks kesehatan (Zheng et al., 2024). Jika sebuah aplikasi mengklasifikasikan hewan yang sakit parah sebagai sehat dengan tingkat keyakinan tinggi, pemilik dapat menunda penanganan yang penting, sehingga mengikis kepercayaan terhadap teknologi ini; transparansi terkait ketidakpastian model karenanya menjadi prasyarat etis bagi dukungan keputusan klinis yang aman (Yu et al., 2025). Untuk memitigasi risiko ini, sistem yang diusulkan mengintegrasikan mekanisme *Confidence Calibration* menggunakan Temperature Scaling, yang menyelaraskan skor kepercayaan model dengan akurasi empirisnya sehingga probabilitas yang ditampilkan benar-benar merefleksikan kemungkinan kebenarannya (Cao et al., 2024). Skor kepercayaan yang terkalibrasi dengan baik memungkinkan sistem menetapkan ambang batas keamanan yang rasional, secara otomatis merekomendasikan

rujukan ke dokter hewan setiap kali tingkat kepercayaan berada di bawah level yang ditentukan, sehingga mengubah AI dari sekadar mesin prediksi menjadi alat manajemen risiko yang bertanggung jawab dan tidak menyesatkan pemilik hewan awam (Du et al., 2025).

Berdasarkan latar belakang di atas, penelitian ini mensintesiskan berbagai tantangan tersebut ke dalam satu sistem terintegrasi yang, sepengetahuan penulis, belum pernah diimplementasikan sebelumnya di Indonesia. Novelti utamanya terletak pada kombinasi arsitektur Two-Stage Lesion-Aware untuk presisi dengan mekanisme Confidence Calibration untuk penjaminan keamanan medis. Penelitian ini diarahkan pada tiga tujuan yaitu yang pertama mengimplementasikan algoritma object-detection untuk lokalisasi lesi (Tahap 1) beserta jaringan klasifikasi untuk analisis penyakit (Tahap 2), yang kedua menerapkan teknik Confidence Calibration untuk meminimalkan calibration error dan menghasilkan metrik ketidakpastian yang andal untuk triase klinis; dan yang ketiga mengukur usability dari alat screening yang dihasilkan, baik bagi pemilik hewan maupun dokter hewan.

2. METODOLOGI

Penelitian ini mengembangkan sistem screening penyakit kulit pada kucing dan anjing berbasis *two-stage lesion-aware pipeline* yang diintegrasikan dengan mekanisme *confidence calibration*. Pengembangan perangkat lunak dilakukan menggunakan pendekatan Agile-Kanban untuk mengakomodasi kebutuhan eksperimen model *machine learning* yang bersifat iterative, sementara arsitektur sistem menggunakan pola client-server dengan pemisahan layanan inferensi AI(*AI microservice*) dari layanan bisnis utama

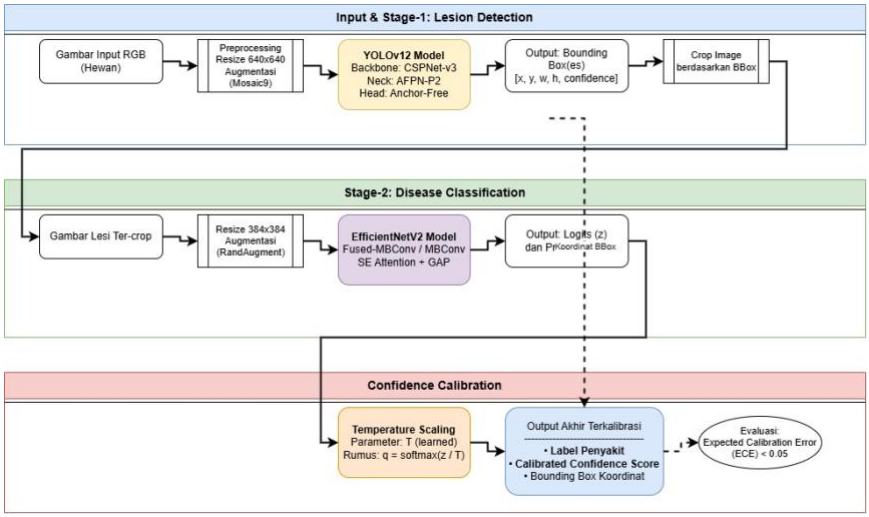
2.1. Dataset

Dataset yang digunakan dalam penelitian ini dihimpun dari tiga sumber publik yang dipilih berdasarkan relevansi penyakit, kualitas anotasi, dan variasi kondisi pengambilan gambar yaitu yangb pertama dataset *Cat Skin Disease* di Kaggle, yang kedua dataset berbasis Roboflow yang dikurasi oleh Nofal Arif dengan anotasi format YOLO (.txt) native, dan yang ketiga dataset *Dog's Skin Diseases* yang dihimpun oleh Youssef Mohammed di Kaggle. Setelah proses deduplikasi dan verifikasi kualitas, dataset yang terkumpul berjumlah total 5.200 gambar (2.800 gambar kucing dan 2.400 gambar anjing), yang dibagi menggunakan skema stratified 80:10:10 menjadi training, validation, dan test set, sehingga proporsi kelas tetap terjaga di seluruh subset. Skema label akhir yang digunakan untuk classifier Tahap-2 terdiri dari tujuh kelas untuk kucing (*Demodicosis, Dermatitis, Flea Allergy, Fungus, Ringworm, Scabies, dan Healthy*) dan enam kelas untuk anjing (*Ringworm, Hypersensitivity, Fungal Infection, Dermatitis, Demodicosis, dan Healthy*). Model klasifikasi yang terpisah dilatih untuk masing-masing spesies agar fitur morfologis yang spesifik terhadap karakteristik kulit dan bulu masing-masing spesies dapat dipelajari secara independen.

2.2 Arsitektur Two-Stage Lesion-Aware Pipeline

Inti dari sistem yang diusulkan adalah sebuah Two-Stage Lesion-Aware Pipeline yang meniru proses diagnostik dua langkah yang dilakukan oleh dokter hewan: pertama mengidentifikasi dan melokalisasi area lesi, kemudian menganalisis area tersebut untuk mengklasifikasikan penyakit. Secara formal, diagnosis otomatis diperlakukan sebagai estimasi probabilitas kondisional $P(Y|X)$, di mana X adalah gambar input dan Y adalah label diagnosis. Sebuah classifier end-to-end tunggal yang mencoba mempelajari pemetaan langsung $f: X \rightarrow$

Y sering kali kurang optimal karena fitur anatomis normal dan noise latar belakang yang dominan mengalihkan perhatian jaringan, sehingga menurunkan spesifisitas (Tekin et al., 2025). Formulasi two-stage sebaliknya memperkenalkan variabel laten *R* yang merepresentasikan Region of Interest (ROI) lesi dan memfaktorkan probabilitas gabungan menggunakan chain rule, $P(Y, R \mid X) = P(Y \mid R, X) \cdot P(R \mid X)$, di mana suku pertama ditangani oleh detektor Tahap-1 dan suku kedua oleh classifier Tahap-2 yang beroperasi pada sub-gambar hasil crop X_R, yang memiliki signal-to-noise ratio jauh lebih tinggi dibanding gambar penuh dan karenanya memfasilitasi konvergensi yang lebih baik pada jaringan klasifikasi (Kannaiya Raja et al., 2025). Gambar 1 mengilustrasikan alur lengkap pipeline yang diusulkan, mulai dari gambar input mentah, deteksi lesi, klasifikasi penyakit, hingga kalibrasi kepercayaan



Gambar 1. Alur Two-Stage Lesion-Aware Pipeline dengan Confidence Calibration.

Tahap 1 (deteksi lesi) menggunakan YOLOv12-Small, anggota keluarga *You Only Look Once* berbasis attention yang memperkenalkan mekanisme Area-Attention (AAttn/ABlock) dan Residual Efficient Layer Aggregation Networks (R-ELAN) untuk menangkap dependensi jarak jauh sambil tetap mempertahankan kecepatan inferensi setara CNN (Tian et al., 2025). Integrasi native FlashAttention mengurangi kompleksitas memori dari operasi attention, dari kuadratik menjadi mendekati linear terhadap panjang sekuens, dengan melakukan tiling pada komputasi attention di dalam SRAM GPU yang cepat (Dao et al., 2022), yang sangat penting untuk inferensi real-time pada gambar medis beresolusi tinggi. Selain regresi bounding-box, head YOLOv12 pada penelitian ini juga menghasilkan mask segmentasi instance melalui cabang berbasis prototipe, yang memungkinkan penggambaran batas lesi secara presisi dan real-time (Khanam & Hussain, 2025).

Tahap 2 (klasifikasi penyakit) menggunakan EfficientNetV2-S, yang diinisialisasi dengan bobot pretrained ImageNet-21K, diterapkan secara eksklusif pada area lesi yang telah dipotong oleh Tahap 1. EfficientNetV2 mengganti konvolusi depthwise-separable pada layer awal dengan blok fused MBConv agar lebih optimal memanfaatkan akselerator modern (Tan & Le, 2021), serta menerapkan progressive learning, di mana resolusi gambar dan kekuatan regularisasi ditingkatkan secara bersamaan seiring berjalannya training, sehingga jaringan dapat mempelajari bentuk lesi global yang kasar terlebih dahulu sebelum menyempurnakan detail tekstur yang halus (Azmoodeh-Kalati et al., 2025). Desain lesion-aware ini berbeda dari klasifikasi berbasis gambar penuh konvensional yang mengandalkan Global Average Pooling

dan sering membuat sinyal lesi kecil larut karena didominasi latar belakang (Nirthika et al., 2022); sebaliknya, membatasi receptive field classifier pada ROI hasil Tahap-1 mengarahkan kapasitas representasional jaringan pada tekstur dan morfologi yang relevan secara patologis (Qian & Yi, 2025).

Tahap 3 (kalibrasi kepercayaan) menerapkan *Temperature Scaling* pada logit mentah yang dihasilkan classifier Tahap-2 sebelum operasi softmax. Dengan vektor logit z, probabilitas terkalibrasi dihitung sebagai $q = softmax(z / T)$, di mana skalar temperature T adalah satu-satunya parameter bebas yang dioptimasi pada validation set. Temperature Scaling merupakan metode kalibrasi post-hoc yang menskalakan ulang kepercayaan tanpa mengubah kelas prediksi model, dan banyak diadopsi karena kesederhanaannya serta sifatnya yang menjaga akurasi klasifikasi sambil meningkatkan keselarasan antara kepercayaan yang diprediksi dan akurasi empiris (Tomani et al., 2022).

2.3 Konfigurasi Pelatihan

Kedua detektor YOLOv12s (masing-masing untuk kucing dan anjing) di-fine-tune dari bobot pretrained COCO menggunakan resolusi input 640×640, batch size 16, dan maksimum 100 epoch dengan early stopping (patience = 20). Augmentasi mosaic (probabilitas 1,0, dinonaktifkan pada 10 epoch terakhir), random scaling (0,9), dan Mixup (0,5) diterapkan untuk meningkatkan generalisasi. Pelatihan dilakukan pada Google Colaboratory Pro menggunakan GPU NVIDIA A100 (VRAM 40 GB); model kucing konvergen setelah 100 epoch (2 jam 54 menit) sedangkan model anjing dilatih selama kurang lebih 261 epoch (3 jam 46 menit) sebelum early stopping.

Classifier EfficientNetV2-S dilatih dengan resolusi input 224×224, batch size 32, dan maksimum 30 epoch dengan early stopping (patience = 7), menggunakan Stochastic Gradient Descent dengan momentum 0,95 dan weight decay 5×10^{-4}. Strategi differential learning-rate digunakan, dengan backbone pretrained diperbarui pada rate yang lebih kecil (0,001) dibandingkan head klasifikasi yang baru diinisialisasi (0,01) untuk menghindari catastrophic forgetting, mengikuti scheduler CosineAnnealingLR. Normalisasi standar ImageNet diterapkan, dan gambar training set diaugmentasi dengan random crop, horizontal flip, dan color jitter, sementara gambar validasi hanya menerima transformasi resize dan center-crop yang deterministik.

2.4. Prosedur Confidence Calibration

Kualitas kalibrasi dikuantifikasi menggunakan Expected Calibration Error (ECE) dan Maximum Calibration Error (MCE). Prediksi dikelompokkan ke dalam 15 bin kepercayaan dengan lebar yang sama; untuk setiap bin B_m, ECE mengagregasi selisih absolut yang dibobot berdasarkan jumlah sampel antara rata-rata kepercayaan prediksi dan akurasi empiris pada bin tersebut, $ECE = \Sigma\ (|B_m|/n) \cdot |acc(B_m) - conf(B_m)|$, sedangkan MCE hanya melaporkan selisih terburuk tunggal, $MCE = max_m\ |acc(B_m) - conf(B_m)|$, yang relevan secara klinis karena menangkap rentang kepercayaan paling menyesatkan tanpa memandang seberapa jarang kemunculannya (Karandikar et al., 2021). Temperature optimal T diperoleh dengan meminimalkan negative log-likelihood dari logit validation set menggunakan optimizer L-BFGS, dimulai dari nilai awal T = 1,5.

2.5. Implementasi Sistem

Pipeline tiga tahap ini di-deploy di belakang backend FastAPI yang mengorkestrasi deteksi lesi, klasifikasi, dan kalibrasi secara berurutan melalui satu endpoint *POST /diagnose*, yang mengembalikan label penyakit prediksi, skor kepercayaan terkalibrasi, kategori tingkat kepercayaan (Tinggi/Sedang/Rendah), dan bounding box lesi ke aplikasi mobile Flutter yang digunakan oleh tiga jenis pengguna yaitu pemilik hewan, dokter hewan, dan administrator. Autentikasi pengguna serta master data yang jarang ditulis (pengguna, dokter hewan, klinik, hewan peliharaan, dan daftar ras) dikelola melalui Firebase Authentication dan Firestore, sementara riwayat diagnosis dan gambar lesi yang sering ditulis dan membutuhkan query relasional dikelola melalui Supabase (PostgreSQL dan Object Storage).

2.6. Metodologi Evaluasi

Detektor Tahap-1 dievaluasi pada test set terpisah (held-out) per spesies menggunakan mAP@0.5, mAP@0.5:0.95, precision, dan recall, yang dihitung otomatis oleh framework Ultralytics. Classifier Tahap-2 dievaluasi menggunakan precision, recall, F1-score, dan akurasi per kelas (melalui classification report scikit-learn), bersama dengan akurasi Top-1 dan Top-5. Kalibrasi dinilai dengan membandingkan ECE dan MCE sebelum dan sesudah Temperature Scaling, baik pada validation set maupun test set independen berjumlah 745 sampel. Waktu respons sistem diukur dengan menginstrumentasi setiap tahap pipeline menggunakan *time.time()* pada Python, dirata-ratakan dari tiga trial dalam kondisi jaringan Wi-Fi lokal yang stabil. Usability dievaluasi melalui dua instrumen yang saling melengkapi kuesioner System Usability Scale (SUS) (sepuluh item, skala Likert 1–5, polaritas positif/negatif bergantian) yang diisi oleh lima responden pemilik hewan setelah melakukan tugas diagnostik terskrip, dengan task success rate, waktu penyelesaian, dan frekuensi jenis error yang dicatat secara bersamaan; serta kuesioner Expert-Based Evaluation (EBE) (skala Likert 1–5) yang diisi oleh tiga dokter hewan praktisi, mencakup empat aspek yaitu kesesuaian pengetahuan (knowledge fit), akurasi diagnostik, kemudahan antarmuka, dan kelayakan sistem secara keseluruhan.

3. HASIL DAN PEMBAHASAN

3.1. Kinerja Deteksi Lesi Tahap-1

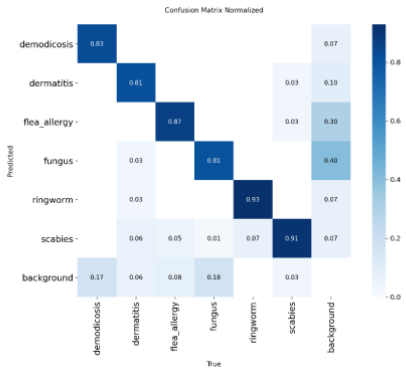
Tabel 1 merangkum kinerja deteksi dari dua model YOLOv12s pada test set independen masing-masing. Model kucing mencapai mAP@0.5 sebesar 0,89 dengan precision 0,88 dan recall 0,84, sedangkan model anjing mencapai mAP@0.5 yang jauh lebih rendah yaitu 0,64, dengan mAP@0.5:0.95 hanya 0,32 dibandingkan 0,61 pada model kucing.

Table 1. Ringkasam kinerja deteksi Tahap-1 (YOLOv12).

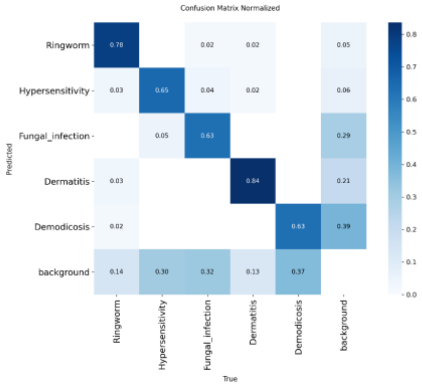
Model	mAP@0.5	mAP@0.5:0.95	Precision	Recall

YOLOv12s – Kucing (6 kelas)	0.89	0.61	0.88	0.84
YOLOv12s – Anjing (5 kelas)	0.64	0.32	0.75	0.64

Kelas Ringworm mencatat kinerja terbaik pada model kucing (recall 0,93, F1 0,94), konsisten dengan jumlah sampelnya yang terbesar dan lesi berbentuk lingkaran alopesik yang khas secara visual, sementara Dermatitis dan Fungus mencatat recall terendah (0,81) akibat tumpang tindih morfologis antara kedua jenis lesi tersebut. Pada model anjing, Dermatitis mencapai recall tertinggi (0,84), sementara Fungal_Infection dan Demodicosis mencatat recall terendah (0,63). Selisih kinerja antar spesies ini dapat diatributkan pada tiga faktor yang saling memperkuat: jumlah training set yang lebih kecil untuk anjing (2.400 vs. 2.800 gambar), variansi morfologi bulu antar-ras yang jauh lebih tinggi, dan durasi pelatihan yang jauh lebih panjang (250 epoch) yang menyebabkan overfitting ringan pada epoch-epoch akhir, sehingga checkpoint terbaik yang digunakan dalam evaluasi diambil dari sekitar epoch 120, bukan epoch terakhir.



Gambar 2. Confusion matri deteksi ternormalisasi – model kucing.



Gambar 3. Confusion matri deteksi ternormalisasi – model anjing

Confusion matrix deteksi (Gambar 2 dan 3) menunjukkan bahwa kebingungan (confusion) dengan latar belakang jauh lebih besar pada model anjing: kelas Demodicosis, Fungal_Infection, dan Hypersensitivity menunjukkan tingkat false-negative dan false-positive sebesar 0,30–0,39 terhadap kelas background, dibandingkan maksimum 0,30 (Flea_allergy) pada model kucing. Hal ini mengindikasikan bahwa variasi visual pada lesi anjing membuat

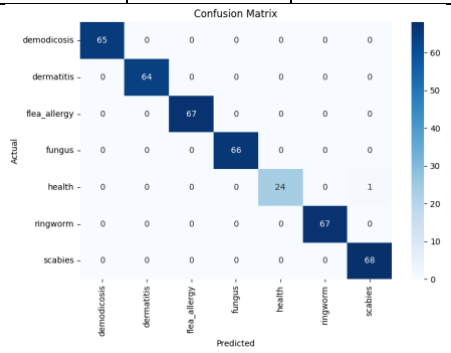
batas antara lesi dan kulit normal jauh lebih sulit diresolusi oleh detektor dibandingkan pada kucing.

3.2. Kinerja Deteksi Lesi Tahap-2

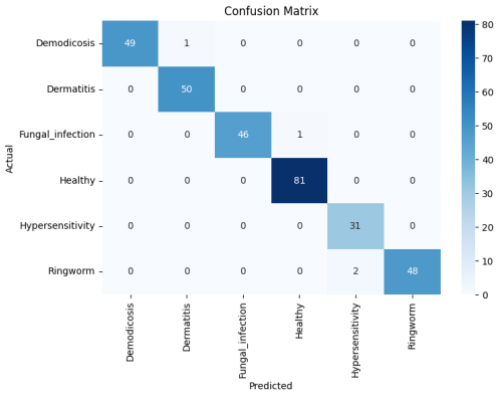
Tabel 2 merangkum hasil klasifikasi EfficientNetV2-S pada Tahap-2. Model kucing memperoleh F1-score weighted-average sebesar 0,9976, dengan hanya satu kesalahan klasifikasi dari 422 sampel evaluasi (satu sampel Healthy diprediksi sebagai Scabies). Model anjing memperoleh F1-score weighted-average sebesar 0,9871 pada 309 sampel, dengan total empat kesalahan klasifikasi, dengan pola paling menonjol berupa kebingungan antara Ringworm dan Hypersensitivity, yang secara klinis masuk akal karena kedua kondisi tersebut dapat menunjukkan kemerahan dan iritasi sebelum pola melingkar khas Ringworm menjadi jelas. Kedua model mencapai akurasi Top-5 sebesar 1,0000, artinya kelas sebenarnya selalu berada di antara lima prediksi teratas yang ditampilkan ke antarmuka pengguna, memberikan margin keamanan tambahan di luar prediksi Top-1.

Table 2. Ringkasan kinerja klasifikasi Tahap-2 (EfficientNetV2-S) (weighted average).

Model	Precision	Recall	F1-Score	Akurasi Top-1	Akurasi Top-5
Kucing (7 kelas, n=422)	0.9977	0.9976	0.9976	0.9976	1.0000
Anjing (6 kelas, n=309)	0.9876	0.9871	0.9871	0.9871	1.0000



Gambar 4. Confusion matrix klasifikasi (jumlah absolut) – model kucing



Gambar 5. Confusion matrix klasifikasi (jumlah absolut) – model anjing

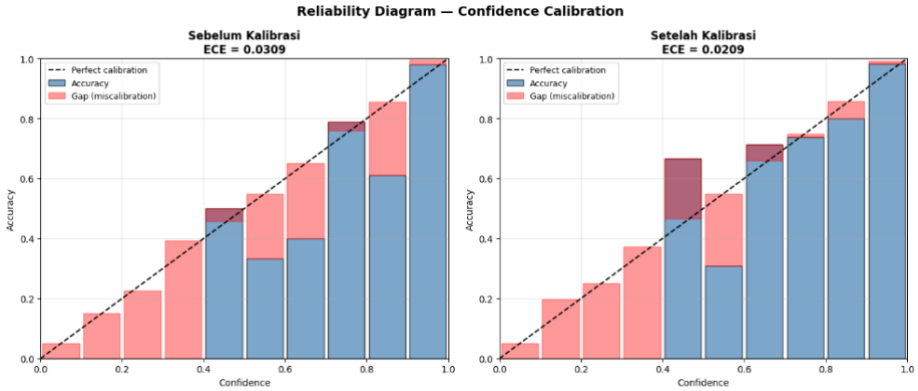
Skor yang hampir sempurna ini memerlukan interpretasi akademik yang hati-hati. Karena akurasi yang dilaporkan identik dengan metrik *valid/top@1* yang tercatat selama pelatihan, evaluasi pada Tabel 2 kemungkinan besar masih dilakukan pada split validasi yang sama dengan yang digunakan untuk pemilihan checkpoint dan early stopping, bukan pada test set yang benar-benar belum pernah dilihat sepanjang seluruh proses pengembangan model. Meskipun data validasi berasal dari pembagian berbasis sumber (source-level split) dan bukan pembagian acak, keterbatasan metodologis ini perlu diungkapkan secara transparan dan evaluasi pada test set yang benar-benar independen direkomendasikan pada penelitian selanjutnya untuk memperkuat validitas klaim kinerja ini. Meskipun demikian, jumlah error absolut yang rendah pada kedua spesies konsisten dengan keunggulan teoretis dari desain two-stage: memisahkan lokalisasi dari klasifikasi memungkinkan jaringan Tahap-2 untuk fokus secara spesifik pada representasi area lesi, sehingga akurasi klasifikasi tetap tinggi meskipun kinerja deteksi Tahap-1 (khususnya untuk anjing) relatif lebih lemah.

3.3. Kinerja Confidence Calibration

Tabel 3 menyajikan hasil kalibrasi sebelum dan sesudah Temperature Scaling. Temperature optimal yang diperoleh melalui optimasi L-BFGS pada validation set adalah T = 1,3174. Pada test set independen (745 sampel yang tidak digunakan selama pencarian temperature), ECE menurun dari 0,0309 menjadi 0,0209, penurunan sebesar 32,4%, sementara MCE menurun dari 0,25 menjadi 0,22, penurunan yang lebih kecil sebesar 12%. Model mencapai akurasi 94,77% pada test set, mengonfirmasi bahwa kalibrasi tidak menurunkan kinerja klasifikasi.

Table 3. Kinerja kalibrasi sebelum dan sesudah Temperature Scaling

Metrik	Sebelum Kalibrasi	Sesudah Kalibrasi	Penurunan
ECE – Validation Set	0.0224	0.0154	31.3%
ECE – Test Set	0.0309	0.0209	32.4%
MCE	0.25	0.22	12%
Temperature Optimal (T)	1.5 (awal)	1.3174	–
Akurasi Test Set	–	94.77%	–



Gambar 6. Reliability diagram sebelum (kiri, ECE = 0,0309) dan sesudah (kanan, ECE = 0,0209) Temperature Scaling.

Nilai temperature yang lebih besar dari 1 mengonfirmasi bahwa model sebelum dikalibrasi bersifat sedikit *underconfident*, yaitu distribusi probabilitas keluarannya lebih rata (flat) dibandingkan akurasi sebenarnya, yang dikonfirmasi oleh reliability diagram pada Gambar 6 melalui bar akurasi biru yang secara sistematis berada di atas garis diagonal pada rentang kepercayaan menengah (0,4–0,8) sebelum kalibrasi. Deviasi *T* yang relatif kecil dari 1,0 mengindikasikan bahwa model dasar sudah cukup terkalibrasi dengan baik dan hanya memerlukan koreksi halus, bukan koreksi drastis. Penurunan MCE yang jauh lebih kecil dibandingkan ECE (12% vs. 32,4%) merupakan karakteristik inheren dari Temperature Scaling, yang mengoptimasi satu parameter global sehingga kurang sensitif terhadap miskalibrasi lokal pada bin kepercayaan tertentu limitasi yang sebagian dimitigasi oleh lapisan aplikasi dengan memicu rekomendasi rujukan otomatis ke dokter hewan setiap kali tingkat kepercayaan berada di bawah ambang batas yang ditentukan.

3.4. Waktu Respons Sistem

Waktu respons end-to-end diukur pada tiga trial menggunakan gambar uji berukuran rata-rata (±800 KB) pada jaringan Wi-Fi lokal yang stabil. Tabel 4 menunjukkan bahwa rata-rata total waktu respons pipeline adalah 4,973 detik, jauh di bawah persyaratan non-fungsional 15 detik. Komputasi AI itu sendiri hanya menyumbang sebagian kecil dari total waktu (deteksi lesi: 0,805 s; klasifikasi: 1,002 s; kalibrasi: 0,157 s), sementara operasi I/O jaringan mengunggah gambar ke Supabase Storage (1,829 s) dan menulis rekam diagnosis ke database (1,179 s) merepresentasikan porsi terbesar dari latensi end-to-end.

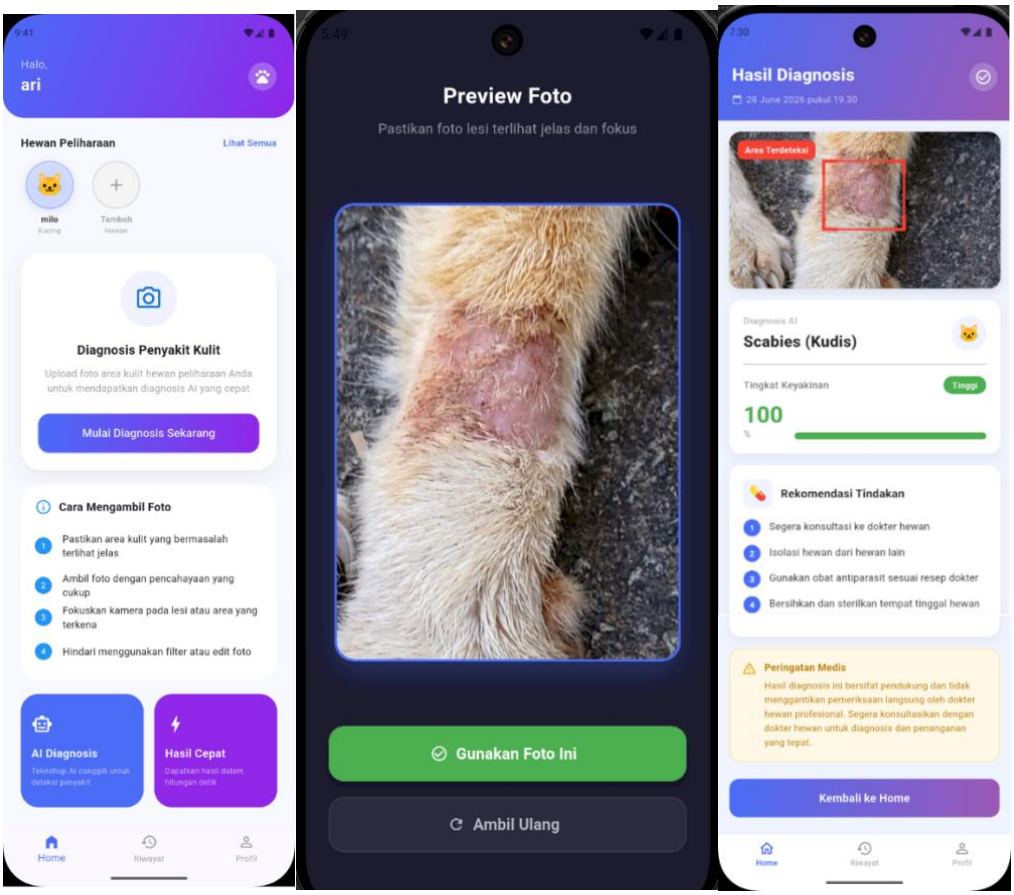
Table 4. Rata-rata waktu respons pipeline pada tiga trial.

Tahap Pipeline	Rata-rata Waktu (s)
Validasi input & penyimpanan gambar sementara	0.001
Tahap 1 – Deteksi lesi (YOLOv12)	0.805
Tahap 2 – Klasifikasi penyakit (EfficientNetV2)	1.002
Tahap 3 – Kalibrasi kepercayaan (Temperature Scaling)	0.157
Unggah gambar ke Supabase Storage	1.829
Penulisan rekam diagnosis ke Supabase DB	1.179
Total waktu respons sistem	4.973

3.5. Implementasi Antarmuka Sistem (Mobile dan Web)

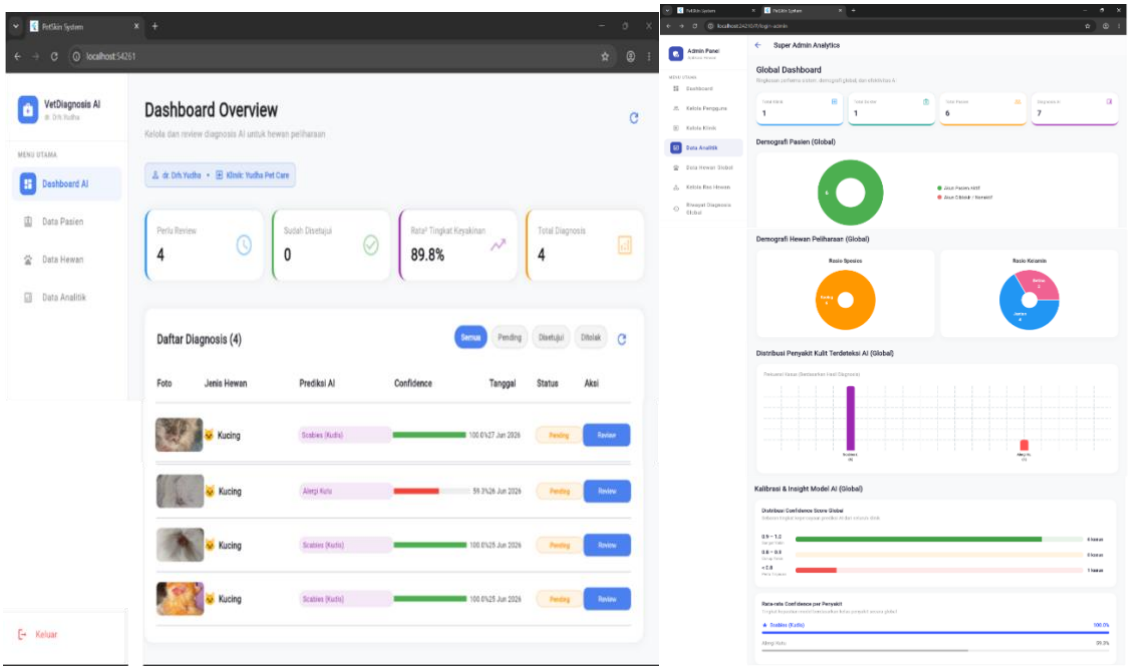
Sebagai bukti akhir validasi sistem di lapangan, pipeline AI yang telah dievaluasi pada Bagian 3.1–3.4 diintegrasikan ke dalam aplikasi mobile Flutter (untuk pemilik hewan) dan aplikasi web (untuk dokter hewan serta administrator), dan diuji langsung oleh responden menggunakan perangkat mobile sungguhan pada sesi pengujian usability dan Expert-Based Evaluation. Gambar 7 menunjukkan tiga antarmuka utama pada sisi pemilik hewan: Halaman Utama yang menyediakan tombol aksi "Mulai Diagnosis Sekarang" beserta panduan teknis pengambilan foto lesi, Halaman Pratinjau Foto sebelum foto dikirim ke pipeline AI, dan Halaman Hasil Diagnosis yang menampilkan label penyakit prediksi beserta skor kepercayaan

terkalibrasi dan kategori tingkat kepercayaan (Tinggi/Sedang/Rendah) sebagai keluaran akhir dari mekanisme confidence calibration yang dijelaskan pada Bagian 2.4.



Gambar 7. Antarmuka aplikasi mobile pemilik hewan — Halaman Utama, Pratinjau Foto, dan Hasil Diagnosis.

Pada sisi dokter hewan dan administrator, sistem menyediakan portal web yang terpisah. Gambar 8 menunjukkan Dashboard Overview pada portal dokter hewan, yang menyajikan ringkasan kasus yang memerlukan review, serta Dashboard Analitik pada Admin Panel, yang menampilkan visualisasi distribusi confidence score dan rata-rata confidence per penyakit — memberikan dokter hewan dan administrator visibilitas langsung terhadap perilaku kalibrasi model yang telah dibahas secara kuantitatif pada Bagian 3.3



Gambar 8. Antarmuka portal web — Dashboard Overview dokter hewan (kiri) dan Dashboard Analitik Admin Panel (kanan).

Pengujian lapangan menggunakan antarmuka nyata ini bukan sekadar mockup rancangan menjadi dasar empiris bagi hasil System Usability Scale dan Expert-Based Evaluation yang disajikan pada Bagian 3.6, sehingga skor usability yang dilaporkan merefleksikan pengalaman aktual pengguna berinteraksi dengan pipeline AI melalui antarmuka produksi, bukan estimasi dari desain konseptual.

3.6. Evaluasi Usability

Tahap akhir dari evaluasi sistem bertujuan untuk memastikan bahwa seluruh modul yang telah dikembangkan mampu beroperasi sesuai dengan spesifikasi kebutuhan fungsional. Pengujian fungsionalitas ini dieksekusi menggunakan metode *Black Box Testing* guna memvalidasi stabilitas interaksi sistem pada seluruh peran pengguna yang terlibat, yang secara spesifik mencakup pemilik hewan peliharaan, Dokter Hewan, serta administrator. Seluruh Skenario pengujian dan statusnya dirangkum pada Tabel 5 hingga Tabel 7

Table 5. Hasil Pengujian Fungsional – Pemilik Hewan

Kode	Skenario Uji	Status
FPH-01	Registrasi akun baru	Berhasil
FPH-02	Registrasi dengan email sudah terdaftar	Berhasil
FPH-03	Login dengan kredensial valid	Berhasil
FPH-04	Login dengan password salah	Berhasil
FPH-05	Pilih foto dari galeri	Berhasil
FPH-06	Preview dan konfirmasi foto	Berhasil

FPH-07	Pilih hewan untuk diagnosis	Berhasil
FPH-08	Upload foto ke server	Berhasil
FPH-09	Upload foto melebihi 10MB	Berhasil
FPH-10	Proses diagnosis dan tampil hasil	Berhasil
FPH-11	Tampil rekomendasi berdasarkan *confidence level*	Berhasil
FPH-12	Tampil disclaimer AI	Berhasil
FPH-13	Tampil riwayat diagnosis	Berhasil
FPH-14	Logout dari aplikasi	Berhasil

Table 6. Hasil Pengujian Fungsional – Pemilik Hewan

Kode	Skenario Uji	Status
FPH-01	Login dokter hewan	Berhasil
FPH-01b	Login dengan akun belum diverifikasi	Berhasil
FPH-03	Tampil dashboard statistik	Berhasil
FPH-04	Tampil detail diagnosis	Berhasil
FPH-05	Approve diagnosis	Berhasil
FPH-06	Reject diagnosis	Berhasil
FPH-07	Reject tanpa mengisi alasan	Berhasil
FPH-08	Koreksi label diagnosis	Berhasil
FPH-09	Logout dokter	Berhasil

Table 7. Hasil Pengujian Fungsional – Pemilik Hewan

Kode	Skenario Uji	Status
FPH-01	Login administrator dengan kredensial valid	Berhasil
FPH-01b	Verifikasi data *real-time* pada dashboard utama	Berhasil
FPH-03	Manajemen data pengguna (aktivasi/nonaktif/hapus)	Berhasil
FPH-04	Logout administrator	Berhasil

Berdasarkan Tabel 5 hingga Tabel 7, seluruh 27 skenario pengujian (14 skenario Pemilik Hewan, 9 skenario Dokter Hewan, dan 4 skenario Administrator) dinyatakan Berhasil (*pass rate* 100%). Cakupan pengujian meliputi alur autentikasi dan otorisasi peran, validasi input

(termasuk penanganan kesalahan seperti email terdaftar dan ukuran unggahan berlebih), proses inti diagnosis berbasis AI beserta tampilan *bounding box* dan rekomendasi berbasis *confidence level*, alur *review*/koreksi diagnosis oleh dokter hewan, hingga manajemen pengguna dan pemantauan statistik oleh administrator. Tidak ditemukan kegagalan fungsional pada seluruh skenario yang diujikan, mengonfirmasi bahwa implementasi sistem telah sesuai dengan spesifikasi kebutuhan fungsional yang ditetapkan pada tahap perancangan.

Setelah pengujian fungsionalitas , responden mengisi kuesioner System Usability Scale (SUS) standar berisi 10 pernyataan dengan skala Likert 1–5, dengan pernyataan bernomor ganjil bersifat positif dan bernomor genap bersifat negatif, yang kemudian dihitung menggunakan rumus standar SUS untuk menghasilkan skor pada rentang 0–100. Tabel 8 menyajikan skor SUS lengkap per responden

Table 8. Skor System Usability Scale (SUS) per responden pemilik hewan

Responden	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Skor SUS
R1	3	4	4	1	5	2	3	3	4	1	70
R2	3	1	4	1	4	1	5	1	5	1	90
R3	5	2	4	3	4	1	4	1	4	2	80
R4	5	1	5	2	5	1	5	1	5	1	95
R5	4	2	5	1	5	1	4	1	5	2	90
Rata-rata											85

Table 9 . Hasil Task-Based Usability Testing

No	Tugas	Berhasil	Berhasil dengan bantuan	Gagagl	Total	Task Success Rate
1	Registrasi akun + diagnosis foto lesi dari galeri (first-time use)	4	1	0	5	80%
2	Menjelaskan confidence score & rekomendasi dari hasil diagnosis	5	0	0	5	100%
	Total Keseluruhan	9	1	0	10	90%

Pengujian usability pada lima responden pemilik hewan menghasilkan task success rate keseluruhan sebesar 90% yang terdapat pada Tabel 8 dan skor SUS rata-rata 85,yang termasuk kategori *Excellent*. Empat dari lima responden mencatat skor pada rentang Excellent (≥ 85), sementara satu responden lainnya mencatat skor 70 (Acceptable), konsisten dengan responden tersebut sebagai satu-satunya yang memerlukan bimbingan navigasi pada tugas pertama. Tidak ada error interpretasi yang tercatat pada responden manapun, mengindikasikan bahwa teks dan label hasil diagnosis cukup jelas bagi pengguna awam; jenis error yang paling umum adalah error navigasi (90% responden) dan error prosedural seperti melewatkan langkah pemilihan hewan sebelum mengaktifkan kamera (80%).

Table 10. Hasil Expert-Based Evaluation (EBE) oleh validator dokter hewan (skala 1–5).

Validator	Kesesuaian Pengetahuan	Akurasi Diagnostik	Antarmuka & Kemudahan	Kelayakan Sistem
Dokter Hewan 1	3.33	3.00	5.00	3.33
Dokter Hewan 2	2.00	1.33	5.00	1.00
Dokter Hewan 3	1.33	1.67	3.33	1.33
Rata-rata	2.22	2.00	4.44	1.89

Expert-Based Evaluation oleh tiga dokter hewan menilai aspek Antarmuka & Kemudahan Penggunaan paling tinggi (rata-rata 4,44, "Sangat Baik"), konsisten dengan temuan SUS di kalangan pemilik hewan, mengonfirmasi kualitas antarmuka sebagai kekuatan yang jelas dari sistem yang dikembangkan. Ketiga aspek klinis lainnya kesesuaian pengetahuan (2,22), akurasi diagnostik (2,00), dan kelayakan sistem (1,89) dinilai jauh lebih rendah, merefleksikan standar evaluatif yang lebih tinggi yang diterapkan oleh profesional medis. Cakupan penyakit saat ini (enam kelas anjing dan tujuh kelas kucing) dinilai belum cukup untuk mewakili spektrum penyakit kulit yang umum dijumpai dalam praktik klinis harian, dan dua dari tiga validator juga menilai akurasi diagnostik rendah pada kasus dengan presentasi visual yang ambigu atau out-of-distribution.

3.7. Pembahasan

Penelitian pertama mengimplementasikan deteksi lesi dan klasifikasi penyakit sebagai pipeline two-stage yang terintegrasi secara fungsional telah tercapai, dibuktikan dengan rata-rata waktu respons end-to-end sebesar 4,973 detik, jauh di bawah ambang batas 15 detik yang dipersyaratkan. Namun, selisih mAP@0.5 sebesar 0,25 poin antara model kucing (0,89) dan anjing (0,64) menunjukkan bahwa kinerja deteksi dalam dermatologi veteriner sangat bergantung pada volume dan representativitas dataset per spesies, bukan hanya pada kapasitas model. Skor klasifikasi Tahap-2 yang hampir sempurna (F1 weighted 0,9976 untuk kucing, 0,9871 untuk anjing) semakin mengonfirmasi keunggulan teoretis dari pemisahan lokalisasi dan klasifikasi,

Penelitian kedua menerapkan confidence calibration untuk triase klinis juga tercapai. Temperature Scaling dengan T = 1,3174 menurunkan ECE test set sebesar 32,4% (dari 0,0309 menjadi 0,0209), melampaui target non-fungsional (ECE < 0,15) lebih dari tujuh kali lipat. Penurunan MCE yang lebih terbatas (12%) merupakan sifat inheren dari parameter scaling

global tunggal dan merepresentasikan limitasi residual yang dapat diatasi pada penelitian selanjutnya melalui metode kalibrasi per-kelas atau per-bin.

Penelitian ketiga mengevaluasi usability sistem menghasilkan hasil yang sangat positif di kalangan pemilik hewan (SUS 85, task success rate 90%, nol error interpretasi) namun gambaran yang lebih moderat di kalangan dokter hewan. Ini bukan sebuah kontradiksi, melainkan merefleksikan perbedaan dimensi evaluatif: SUS mengukur usability antarmuka, sementara EBE mengevaluasi kelayakan klinis dari sudut pandang profesional. Sesuai dengan lingkup yang ditetapkan dalam penelitian ini, sistem dimaksudkan secara ketat sebagai instrumen screening awal, bukan pengganti diagnosis klinis yang komprehensif; skor EBE yang lebih rendah pada aspek klinis karenanya berfungsi sebagai landasan berbasis bukti untuk pengembangan selanjutnya khususnya memperluas cakupan kelas penyakit dan memperkuat validasi pada test set independen bukan sebagai indikasi kegagalan sistem. Secara keseluruhan, ketiga untaian evaluasi (kinerja AI, kalibrasi, dan usability) saling memperkuat dan mendukung posisi sistem sebagai alat screening lini pertama yang responsif dan terkalibrasi dengan baik, yang lingkup klinisnya perlu diperluas sebelum diterapkan secara lebih luas.

4. CONCLUSION

Berdasarkan hasil implementasi, pengujian, dan pembahasan yang telah dilakukan, dapat ditarik kesimpulan sebagai berikut:

1. Penelitian ini berhasil mengimplementasikan Two-Stage Lesion-Aware Pipeline yang memisahkan lokalisasi lesi dari klasifikasi penyakit pada gambar kulit kucing dan anjing. Detektor YOLOv12 pada Tahap-1 mencapai mAP@0.5 sebesar 0,89 (kucing) dan 0,64 (anjing), sementara classifier EfficientNetV2 pada Tahap-2 mencapai F1-score weighted sebesar 0,9976 (kucing) dan 0,9871 (anjing); pipeline terintegrasi, yang dilayani melalui backend FastAPI, menghasilkan rata-rata waktu respons end-to-end sebesar 4,973 detik, jauh di bawah persyaratan 15 detik.
2. Penerapan Kalibrasi Temperature Scaling dengan T optimal = 1,3174 menurunkan Expected Calibration Error pada test set 800ndependent sebesar 32,4% (dari 0,0309 menjadi 0,0209), melampaui target non-fungsional lebih dari tujuh kali lipat dan mengonfirmasi bahwa skor kepercayaan yang ditampilkan cukup andal untuk mendukung rekomendasi triase otomatis.
3. Pengujian usability menghasilkan skor SUS rata-rata 85 (Excellent) dengan task success rate 90% di kalangan pemilik hewan, sementara Expert-Based Evaluation oleh dokter hewan menilai antarmuka tinggi (4,44/5,00) namun mengidentifikasi akurasi diagnostik, kesesuaian pengetahuan, dan cakupan kelas penyakit sebagai aspek yang masih memerlukan pengembangan lebih lanjut sebelum sistem dapat dianggap layak untuk penggunaan klinis profesional.

Penelitian mendatang direkomendasikan untuk berfokus pada perluasan volume dan diversitas *dataset* pelatihan terutama pada representasi citra anjing, guna memitigasi disparitas performa deteksi antarspesies. Evaluasi lanjutan terhadap model klasifikasi (Tahap 2) menggunakan *test set* yang sepenuhnya independen mutlak diperlukan untuk memvalidasi ketahanan model secara empiris. Dari perspektif klinis, penambahan spektrum kelas penyakit disarankan untuk meningkatkan akseptabilitas sistem bagi praktisi veteriner. Selain itu, eksplorasi metode *confidence calibration* spesifik berbasis kelas guna melengkapi

Temperature Scaling dapat diuji untuk menekan angka *Maximum Calibration Error* (MCE). Terakhir, perluasan demografi dan wilayah pengujian *usability* pada berbagai klinik hewan sangat direkomendasikan demi memperkuat generalisasi temuan secara operasional

5. REFERENCES

Abudukelimu, H., Gao, Y., Abulizi, A., Musideke, M., Wu, S., Wang, M., Aizizi, M., Yehaiya, G., & Abudukelimu, M. (2025). DVF-YOLO-Seg: A two-stage breast mass segmentation model with enhanced feature extraction and small lesion detection. *DIGITAL HEALTH*, *11*. https://doi.org/10.1177/20552076251374192

Azmoodeh-Kalati, M., Shabani, H., Maghareh, M. S., Barzegar, Z., & Lashgari, R. (2025). Leveraging an ensemble of EfficientNetV1 and EfficientNetV2 models for classification and interpretation of breast cancer histopathology images. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-06853-6

Bard, A. M., Hinchliffe, S., Chan, K. W., Buller, H., & Reyher, K. K. (2023). 'I Believe What I'm Saying More Than the Test': The Complicated Place of Rapid, Point-of-Care Tests in Veterinary Diagnostic Practice. *Antibiotics*, *12*(5). https://doi.org/10.3390/antibiotics12050804

Cao, Z., Li, Y., Kim, D. H., & Shin, B. S. (2024). Deep Neural Network Confidence Calibration from Stochastic Weight Averaging. *Electronics (Switzerland)*, *13*(3). https://doi.org/10.3390/electronics13030503

Dao, T., Fu, D. Y., Ermon, S., Rudra, A., & Ré, C. (2022). *FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness*. https://doi.org/10.48550/arXiv.2205.14135

Du, Y., Wang, Y., Kong, M., Liang, T., Long, Q., Chen, B., & Zhu, Q. (2025). *Confidence Calibration for Multimodal LLMs: An Empirical Study through Medical VQA*. https://doi.org/10.1007/978-3-032-04978-0_9

Kannaiya Raja, N., Venkata Naga Ramesh, J., Yousef ABaker El-Ebiary, T., Muniyandy, E., Konda Reddy, N., Ravi Kumar, V., Devarasetty, P., & Associate Professor, S. (2025). Pose Estimation of Spacecraft Using Dual Transformers and Efficient Bayesian Hyperparameter Optimization. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 16, Number 4). https://doi.org/10.14569/IJACSA.2025.0160463

Karandikar, A., Cain, N., Tran, D., Lakshminarayanan, B., Shlens, J., Mozer, M. C., & Roelofs, B. (2021). *Soft Calibration Objectives for Neural Networks*. https://doi.org/10.48550/arXiv.2108.00106

Khanam, R., & Hussain, M. (2025). *A Review of YOLOv12: Attention-Based Enhancements vs. Previous Versions*. https://doi.org/10.48550/arXiv.2504.11995

Nirthika, R., Manivannan, S., Ramanan, A., & Wang, R. (2022). Pooling in convolutional neural networks for medical image analysis: a survey and an empirical study. In *Neural*

Computing and Applications (Vol. 34, Number 7, pp. 5321–5347). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s00521-022-06953-8

Qian, W., & Yi, Z. (2025). MedVisioneer: A Lesion-Aware Deep Vision Framework for Enhanced Diagnostic Assistance in CT Image Recognition and Adaptation. *IEEE Access*. https://doi.org/10.1109/ACCESS.2025.3648493

Tan, M., & Le, Q. V. (2021). *EfficientNetV2: Smaller Models and Faster Training*. https://doi.org/10.48550/arXiv.2104.00298

Tekin, H., Kılıç, Ş., & Doğan, Y. (2025). DiagNeXt: A Two-Stage Attention-Guided ConvNeXt Framework for Kidney Pathology Segmentation and Classification. *Journal of Imaging*, *11*(12). https://doi.org/10.3390/jimaging11120433

Tian, Y., Ye, Q., & Doermann, D. (2025). *YOLOv12: Attention-Centric Real-Time Object Detectors*. http://arxiv.org/abs/2502.12524

Tomani, C., Cremers, D., & Buettner, F. (2022). *Parameterized Temperature Scaling for Boosting the Expressive Power in Post-Hoc Uncertainty Calibration*. https://doi.org/10.48550/arXiv.2102.12182

Yu, Y., Gomez-Cabello, C. A., Haider, S. A., Genovese, A., Prabha, S., Trabilsy, M., Collaco, B. G., Wood, N. G., Bagaria, S., Tao, C., & Forte, A. J. (2025). Enhancing Clinician Trust in AI Diagnostics: A Dynamic Framework for Confidence Calibration and Transparency. *Diagnostics*, *15*(17). https://doi.org/10.3390/diagnostics15172204

Zaki, F. H. M., Airin, C. M., Nururrozi, A., Yanuartono, & Indarjulianto, S. (2021). A review: the prevalence of dermatophytosis on cats in Indonesia and Turkey. *BIO Web of Conferences*, *33*. https://doi.org/10.1051/bioconf/20213306004

Zheng, Y., Qiu, Y., Che, H., Chen, H., Zheng, W.-S., & Wang, R. (2024). *Deep Model Reference: Simple yet Effective Confidence Estimation for Image Classification*. https://doi.org/10.1007/978-3-031-72117-5_17