Pengembangan *Model Hybrid Efficientnet–Vision Transformer* Untuk Diagnosis Penyakit Gigi-Mulut Berbasis Citra

Juvenus[1], Aloysius Gonzaga Pradnya Sidhawara[2], Patricia Ardanari [3]
Program Studi Informatika, Fakultas Teknologi Industri, Universitas Atma Jaya Yogyakarta
Jl. Babarsari 43, Sleman 55281, Daerah Istimewa Yogyakarta, Indonesia
Email: [1]210711068@students.uajy.ac.id, [2]Aloysius.gonzaga@uajyc.ac.id, [3]patricia.ardanari@uajy.ac.id

Abstract. *Oral and dental diseases are health problems that require early detection to prevent further complications, while manual image-based diagnosis remains prone to subjectivity and interpretation errors. This study aims to design and develop a deep learning model based on a hybrid EfficientNet–Vision Transformer approach to accurately and consistently classify images of oral and dental diseases. The hybrid model employs EfficientNet as a local feature extractor and a Vision Transformer to capture global contextual information. The dataset consists of six disease classes, each containing 1,800 images. Training was conducted using the Adam optimizer with a learning rate of 0.0001, early stopping, and CutMix data augmentation, with EfficientNet-B3 and a standalone Vision Transformer used as baseline comparators. Evaluation results demonstrate that the hybrid model achieves the highest accuracy of 93.69% with stable and well-balanced performance, indicating its potential as a diagnostic decision support system, despite remaining limitations related to dataset size and distribution.*
Keywords: *Oral and Dental Diseases, Image Classification, EfficientNet, Vision Transformer, CutMix Augmentation*

Abstrak. *Penyakit gigi dan mulut merupakan masalah kesehatan yang memerlukan deteksi dini untuk mencegah komplikasi lebih lanjut, sementara diagnosis berbasis citra secara manual masih rentan terhadap subjektivitas dan kesalahan interpretasi. Penelitian ini bertujuan untuk merancang dan mengembangkan model deep learning berbasis pendekatan hybrid EfficientNet–Vision Transformer guna melakukan klasifikasi citra penyakit gigi dan mulut secara akurat dan konsisten. Model hybrid memanfaatkan EfficientNet sebagai ekstraktor fitur lokal dan Vision Transformer untuk menangkap konteks global. Dataset terdiri dari enam kelas penyakit dengan masing-masing 1.800 citra. Pelatihan dilakukan menggunakan optimizer Adam, learning rate 0,0001, early stopping, serta teknik data augmentation CutMix, dengan EfficientNet-B3 dan Vision Transformer tunggal sebagai pembanding. Hasil evaluasi menunjukkan bahwa model hybrid mencapai akurasi tertinggi sebesar 93,69% dengan performa yang stabil dan seimbang, sehingga berpotensi digunakan sebagai sistem pendukung diagnosis, meskipun masih memiliki keterbatasan pada jumlah dan distribusi dataset.*
Kata Kunci: *Penyakit Gigi dan Mulut, Klasifikasi Citra, EfficientNet, Vision Transformer, CutMix Augmentation*

1. Pendahuluan

Penyakit gigi dan mulut merupakan salah satu permasalahan kesehatan yang paling umum di dunia. Laporan World Health Organization (WHO) menyebutkan bahwa sekitar 3,5 miliar orang secara global mengalami penyakit gigi dan mulut. Di Indonesia, Survei Kesehatan Indonesia (SKI) tahun 2023 menunjukkan bahwa 56,9% penduduk berusia di atas tiga tahun memiliki masalah kesehatan gigi dan mulut [1]. Rendahnya kesadaran pemeriksaan dini serta tingginya biaya perawatan menjadi faktor utama keterlambatan penanganan. Kondisi ini diperparah oleh keterbatasan jumlah dan distribusi dokter gigi di Indonesia yang masih berada di bawah standar rasio ideal WHO, khususnya di luar Pulau Jawa [2]. Oleh karena itu, diperlukan solusi alternatif yang mampu mendukung proses diagnosis awal secara lebih merata dan mudah diakses.

Perkembangan Artificial Intelligence (AI), khususnya *deep learning*, telah menunjukkan potensi besar dalam bidang kedokteran digital, termasuk analisis dan klasifikasi citra medis. *Convolutional Neural Networks* (CNN) merupakan pendekatan utama dalam pemrosesan citra dan telah banyak digunakan untuk mendeteksi berbagai penyakit medis. Salah satu arsitektur CNN yang unggul adalah *EfficientNet*, yang mampu menyeimbangkan akurasi dan efisiensi komputasi melalui pendekatan *compound scaling* [3]. Di sisi lain, *Vision Transformer (ViT)* menawarkan pendekatan baru dalam analisis citra dengan memanfaatkan mekanisme *self-attention* untuk menangkap hubungan global antar fitur, meskipun performanya cenderung bergantung pada ketersediaan data dalam jumlah besar [4].

Mengatasi keterbatasan masing-masing arsitektur, penelitian ini mengusulkan pendekatan *hybrid* yang mengombinasikan *EfficientNet* sebagai ekstraktor fitur lokal dan *Vision Transformer* sebagai pemodel konteks global. Selain itu, teknik augmentasi *CutMix* diterapkan untuk meningkatkan kemampuan generalisasi model pada *dataset* medis yang terbatas [5]. Penelitian ini berfokus pada pembuatan dan evaluasi *model hybrid* EfficientNet–Vision Transformer dalam melakukan klasifikasi citra penyakit gigi dan mulut berbasis enam kelas penyakit, serta implementasinya ke dalam aplikasi berbasis *Streamlit* sebagai sistem pendukung diagnosis awal yang praktis dan mudah digunakan.

2. Tinjauan Pustaka

Bagian ini akan menjelaskan tentang penelitian terdahulu dengan topik-topik yang berhubungan dengan model yang dirancang. Penelitian oleh Wahyuningsih, Nugraha, dan Dwiyansaputra mengkaji klasifikasi penyakit karies melalui citra digital dengan membandingkan arsitektur *Efficientnet*-B0 dan beberapa arsitektur *Convolutional Neural Network* (CNN) lainnya. Penelitian ini menggunakan *dataset* awal sebanyak 1.554 citra yang kemudian diaugmentasi menjadi 6.348 citra, dengan pembagian data *training, validation,* dan *testing* menggunakan rasio 70:15:15. Hasil evaluasi menunjukkan bahwa *Efficientnet*-B0 mampu menghasilkan akurasi yang lebih baik dibandingkan arsitektur CNN lainnya, sehingga dinilai efisien dalam mendeteksi penyakit gigi berbasis citra. Namun, penelitian ini masih menggunakan arsitektur CNN tunggal yang berfokus pada ekstraksi fitur lokal, sehingga kemampuan *model* dalam menangkap hubungan global antar bagian citra penyakit gigi belum dieksplorasi lebih lanjut [6].

Penelitian Lavenia, Ramdani, dan Hoeronis berfokus pada klasifikasi penyakit *pulpitis* menggunakan citra radiografi periapikal dengan metode *Convolutional Neural Network* (CNN). Pada penelitian ini digunakan *Dataset* total 1.000 citra yang dibagi menjadi dua kelas, yaitu *pulpitis* dan normal. Hasil penelitian menunjukkan bahwa penggunaan *optimizer* RMSPROP dan *epoch* 50 mampu menghasilkan akurasi 98,75%. Temuan tersebut menunjukkan bahwa CNN dapat menjadi metode yang efektif untuk membantu diagnosis pulpitis secara otomatis melalui gambar radiografi. Namun, penelitian ini hanya berfokus pada penggunaan CNN murni dan gambar radiografi, dan belum mempertimbangkan metode arsitektur lanjutan yang dapat memodelkan hubungan global pada gambar medis secara lebih komprehensif [7].

Penelitian Andika Nur Pratama mengenai klasifikasi penyakit kulit menggunakan arsitektur *Efficientnet*-B2 juga menunjukkan hasil yang menjanjikan dalam bidang citra medis. Penelitian ini bertujuan untuk mengatasi keterbatasan diagnosis manual yang memerlukan waktu lama dan berpotensi menyebabkan kesalahan. Hasil evaluasi menunjukkan bahwa penggunaan *Efficientnet-B2* memiliki akurasi 84,0. Hasil ini menunjukkan bahwa *Efficientnet-B2* memiliki kinerja yang cukup baik dalam klasifikasi gambar medis. Namun, pada penelitian ini hanya menggunakan model CNN Tunggal tanpa augmentasi lanjutan sehingga belum bisa meningkatkan generalisasi model [8].

Penelitian oleh Reyvan Revolusioner Ar, Agusriyati, dan Sumiarni Moka membahas penggunaan Vision Transformer untuk membantu proses skrining dini *stunting* berdasarkan gambar tubuh balita. Penggunaan *model Vision Transformer* (ViT) bertujuan untuk mengidentifikasi hubungan antar bagian citra tubuh secara global. Hasil pengujian menunjukkan bahwa model ViT mampu mencapai akurasi sebesar 98%. Ini menunjukkan bahwa metode ini memiliki kinerja yang sangat baik dalam tugas-tugas yang memerlukan deteksi dini *stunting*

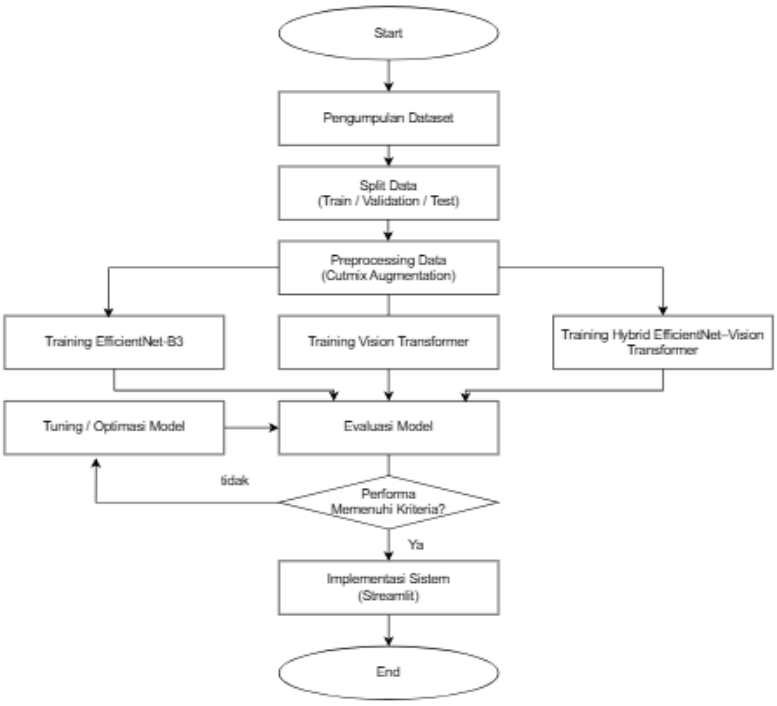
berdasarkan gambar. Namun, penelitian ini masih memiliki keterbatasan karena masih bergantung pada kualitas dan resolusi citra yang baik supaya bisa menghasilkan klasifikasi yang akurat [9].

Penelitian oleh Irvan Yudistiansyah mempelajari penggunaan *Vision transformer* untuk klasifikasi penyakit mata berdasarkan gambar *fundus*. Penelitian ini menunjukkan bahwa penggunaan model *Vision Transformer* dapat memodelkan struktur global dan pola visual kompleks pada gambar *fundus* mata. Hasil pengujian menunjukkan bahwa *model* yang dikembangkan memiliki akurasi sebesar 91,38%. Hal ini menunjukkan bahwa penggunaan model *Vision Transformer* berfungsi dengan baik dalam klasifikasi penyakit mata. Namun, metode ini berpotensi mengalami penurunan performa karena proses pelatihan memerlukan data yang banyak [10].

CNN (EfficientNet) dan Vision Transformer (ViT) telah menunjukkan keberhasilan dalam klasifikasi citra medis secara terpisah, tetapi terdapat tiga celah utama dalam domain penyakit gigi dan mulut, yakni terbatasnya integrasi model *hybrid*, minimnya eksplorasi teknik augmentasi, serta kurangnya implementasi model ke dalam sistem aplikasi praktis yang dapat digunakan langsung sebagai alat bantu diagnosis di lapangan. Oleh karena itu, penelitian ini mengembangkan model *hybrid* EfficientNet–Vision Transformer yang akan dibandingkan performanya dengan model-model lain pada kasus penyakit gigi dan mulut. Penelitian ini juga menerapkan teknik *CutMix Augmentation* guna mengatasi keterbatasan *dataset* medis dan meningkatkan ketahanan model dibandingkan augmentasi konvensional. Selain itu, penelitian ini mengimplementasikan model terbaik ke dalam aplikasi berbasis Streamlit, mengubah hasil riset teoritis menjadi sistem diagnosis praktis yang siap pakai di sektor kesehatan gigi.

3. Metodologi Penelitian

Tahapan yang digunakan dalam penelitian ini yaitu: Pengumpulan data, *preprocessing* data, perancangan dan pengembangan model, pelatihan model, evaluasi model, dan implementasi model (lihat Gambar 1).



Gambar 1. *Flowchart* Alur Penelitian

Tahapan pertama yaitu tahapan pengumpulan data. *Dataset* yang digunakan berasal dari *Kaggle* dengan judul "*Oral Disease*" [11]. Enam kelas yang diklasifikasikan dari *dataset* adalah *Caries, Calculus, Gingivitis, Tooth Discolouration, Mouth Ulcers, dan Hypodontia*. Pada tahapan

ini juga dilakukan analisis *dataset*, seperti menghitung jumlah data, distribusi kelas, dan kualitas citra (lihat Tabel 1). Setelah itu, *dataset* tersebut dibagi menjadi tiga *folder* yaitu *folder training, folder validation,* dan *folder testing.* Untuk pembagiannya menjadi 70:15:15.

Tabel 1. Jumlah Citra Masing-Masing Kelas pada *Dataset* Sebelum Augmentasi

Kelas	Folder Training	Folder Validation	Folder Testing	Total
Calculus	908	194	194	1296
Caries	1273	273	273	1819
Gingivitis	1644	353	353	2350
Hypodontia	877	187	187	1251
Mouth Ulcers	1779	381	381	2541
Tooth Discolouration	1284	275	275	1834

Tahap selanjutnya adalah tahap *preprocessing data.* Ukuran citra diubah menjadi 224 x 224 piksel supaya ukuran semua citra seragam. Setelah ukuran citra seragam, dilakukan teknik augmentasi dasar dan *Cutmix Augmentation.* Augmentasi dasar yang dilakukan meliputi rotasi citra, pergeseran horizontal dan vertikal, perubahan skala (zoom), penyesuaian tingkat kecerahan, serta *horizontal flipping*. Kemudian, untuk menjaga keseimbangan jumlah data antar kelas dalam proses pelatihan, jumlah data latih pada setiap kelas dibatasi menjadi maksimal 1.800 citra dengan mengombinasikan citra asli dan hasil augmentasi dasar (lihat Tabel 2). *CutMix Augmentation* diterapkan ketika proses pelatihan model dijalankan per *batch*. Penerapan *CutMix Augmentation* bertujuan untuk memperkaya variasi data latih dengan cara memotong menjadi beberapa bagian dan mengombinasikan potongan-potongan citra per kelas [5].

Tabel 2. Jumlah Citra Masing-Masing Kelas pada Training Set Sesudah Augmentasi

Kelas	Jumlah Gambar Asli	Jumlah Gambar Augmentasi	Total Setelah Augmentasi
Calculus	908	892	1800
Caries	1273	527	1800
Gingivitis	1644	156	1800
Hypodontia	877	923	1800
Mouth Ulcers	1779	21	1800
Tooth Discolouration	1284	516	1800

Tahap ketiga pada penelitian ini adalah perancangan dan pengembangan arsitektur *model hybrid EfficientNet–Vision Transformer* dengan alur pemrosesan bertahap (sequential *hybrid*). Pada pendekatan ini, *Vision Transformer* digunakan terlebih dahulu untuk memodelkan hubungan global dan konteks spasial citra melalui mekanisme *self-attention*, kemudian representasi fitur yang dihasilkan diteruskan ke *EfficientNet* untuk mengekstraksi fitur lokal yang lebih detail seperti tekstur dan pola visual spesifik. Pendekatan ini tidak menerapkan penggabungan fitur secara *paralel*, melainkan memanfaatkan keunggulan masing-masing arsitektur secara berurutan guna menghasilkan representasi fitur yang lebih kaya dan diskriminatif. Selain perancangan arsitektur, tahap ini juga mencakup penentuan struktur jaringan, konfigurasi parameter pelatihan, seperti *learning rate,* dan *optimizer*, serta mekanisme evaluasi model selama proses pelatihan.

Tahap keempat penelitian ini adalah pelatihan *model hybrid EfficientNet–Vision Transformer* melalui pendekatan *supervised learning* untuk mengoptimalkan parameter model dalam klasifikasi penyakit gigi dan mulut. Model dilatih menggunakan data yang telah melalui proses *preprocessing* dan *augmentasi*, dengan jumlah maksimum 1.800 citra per kelas. Proses pelatihan dilakukan secara *end-to-end* dengan alur sekuensial, di mana *Vision Transformer* terlebih dahulu memodelkan konteks global citra, kemudian *EfficientNet* memperkuat ekstraksi fitur lokal yang relevan. Optimasi model menggunakan *optimizer AdamW* dengan *learning rate* 1×10^{-4} dan fungsi kerugian *Cross Entropy Loss*. Pelatihan dilakukan selama 30 *epoch* dengan pemantauan nilai *loss* dan akurasi pada data pelatihan serta validasi, dan model dengan akurasi validasi tertinggi disimpan sebagai model terbaik untuk tahap evaluasi akhir.

Evaluasi model dilakukan setelah proses pelatihan selesai dengan menggunakan data uji *(test set)* yang tidak terlibat dalam tahap pelatihan maupun validasi guna menilai kemampuan generalisasi model dalam mengklasifikasikan citra penyakit gigi dan mulut. Model terbaik yang diperoleh selama pelatihan menghasilkan keluaran berupa probabilitas untuk setiap kelas, yang kemudian dikonversi menjadi label prediksi berdasarkan nilai probabilitas tertinggi. Performa model dievaluasi menggunakan metrik *accuracy, precision, recall,* dan *F1-score* untuk setiap kelas, serta nilai rata-rata *macro* dan *weighted* guna memberikan gambaran kinerja secara keseluruhan. Selain itu, *confusion matrix* digunakan untuk menganalisis pola kesalahan klasifikasi antar kelas, dan metrik *ROC–AUC* dengan pendekatan *One-vs-Rest* diterapkan untuk menilai kemampuan pemisahan kelas. Hasil evaluasi pada data uji kemudian dibandingkan dengan data pelatihan dan validasi untuk memastikan model tidak mengalami *overfitting* dan memiliki performa yang stabil.

Tahap terakhir adalah implementasi sistem, yaitu *model* terbaik hasil pelatihan dan evaluasi diintegrasikan ke dalam aplikasi berbasis *Streamlit*. Implementasi ini bertujuan untuk menunjukkan bahwa *model* yang dikembangkan tidak hanya memiliki performa yang baik secara teoritis, tetapi juga dapat digunakan secara praktis sebagai sistem klasifikasi citra penyakit gigi dan mulut berbasis antarmuka pengguna.

4. Hasil dan Diskusi

Hasil diperoleh dari proses pelatihan dan pengujian *model hybrid EfficientNet–Vision Transformer* menggunakan *dataset* citra penyakit gigi dan mulut yang telah melalui tahap pra-pemrosesan, augmentasi, dan *feature engineering*. Evaluasi performa model dilakukan menggunakan beberapa metrik evaluasi yang relevan untuk menilai kinerja model secara objektif dan terukur.

4.1 Hasil Evaluasi Kuantitatif

Berdasarkan *classification report* pada data uji, *model hybrid EfficientNet–Vision Transformer* mencapai akurasi keseluruhan sebesar 93,69% dengan nilai F1-score berbobot (*weighted*) sebesar 0,9387, yang menunjukkan performa klasifikasi yang tinggi dan stabil. Kelas *Mouth Ulcers* dan *Tooth Discolouration* menunjukkan performa terbaik dengan nilai *precision* dan *F1-score* mendekati atau mencapai 1,00, menandakan kemampuan model dalam mengenali pola visual kedua kelas tersebut secara konsisten. Kelas *Hypodontia* dan *Caries* juga memperoleh nilai F1-score yang sangat tinggi, masing-masing sebesar 0,9866 dan 0,9603. Sementara itu, kelas *Calculus* memiliki performa relatif lebih rendah dengan F1-score sebesar 0,7840, yang mengindikasikan adanya kesulitan model dalam membedakan karakteristik visual *Calculus* dari kelas lain, khususnya *Gingivitis*, yang memiliki kemiripan tekstur. Nilai *macro average* dan *weighted average* yang tinggi menunjukkan bahwa model memiliki performa yang seimbang pada sebagian besar kelas meskipun terdapat variasi distribusi data.

Tabel 3. Classification Report Model *Hybrid*

Kelas	Precision	Recall	F1-Score	Support
Calculus	0.7198	0.8608	0.7840	194
Caries	0.9922	0.9304	0.9603	273
Gingivitis	0.8947	0.8669	0.8806	353
Hypodontia	0.9892	0.9840	0.9866	187
Mouth Ulcers	1.0000	1.0000	1.0000	381
Tooth Discolouration	1.0000	0.9673	0.9834	275
Accuracy			0.9369	1663
Macro AVG	0.9327	0.9349	0.9325	1663
Weighted AVG	0.9425	0.9369	0.9387	1663

Berdasarkan hasil pengujian pada data uji yang terdiri dari 1.663 citra, model EfficientNet-B3 mencapai akurasi sebesar 93,08% dengan nilai F1-score berbobot sebesar 0,9316. Nilai ROC-AUC One-vs-Rest (macro) sebesar 0,9929 menunjukkan bahwa model memiliki kemampuan diskriminatif yang sangat tinggi dalam membedakan setiap kelas penyakit

gigi dan mulut. Secara per kelas, model menunjukkan performa yang sangat baik pada sebagian besar kategori, dengan kelas *Mouth Ulcer* dan *Tooth Discolouration* memperoleh nilai F1-score di atas 0,99, serta kelas *Hypodontia* mencapai F1-score sebesar 0,9892. Kelas *Caries* dan *Gingivitis* juga menunjukkan keseimbangan yang baik antara precision dan recall dengan nilai F1-score masing-masing sebesar 0,9526 dan 0,8631. Sementara itu, kelas *Calculus* memperoleh F1-score sebesar 0,7607, yang meskipun lebih rendah dibandingkan kelas lainnya, masih menunjukkan performa yang relatif memadai dalam klasifikasi citra penyakit gigi dan mulut.

Tabel 4. Classification report Model EfficientNet

Kelas	Precision	Recall	F1-Score	Support
Calculus	0.7438	0.7784	0.7607	194
Caries	0.9882	0.9194	0.9526	273
Gingivitis	0.8512	0.8754	0.8631	353
Hypodontia	0.9946	0.9840	0.9892	187
Mouth Ulcers	0.9896	0.9974	0.9935	381
Tooth Discolouration	0.9964	0.9927	0.9945	275
Accuracy			0.9308	1663
Macro AVG	0.9273	0.9245	0.9256	1663
Weighted AVG	0.9330	0.9308	0.9316	1663

Berdasarkan hasil pengujian pada data uji yang terdiri dari 1.663 citra, model Vision Transformer mencapai akurasi sebesar 85,81% dengan nilai F1-score berbobot sebesar 0,8631. Nilai ROC-AUC One-vs-Rest (OvR) sebesar 0,9803 menunjukkan bahwa model memiliki kemampuan diskriminatif yang baik dalam membedakan kelas penyakit gigi dan mulut, meskipun masih berada di bawah performa model *hybrid* dan EfficientNet-B3.

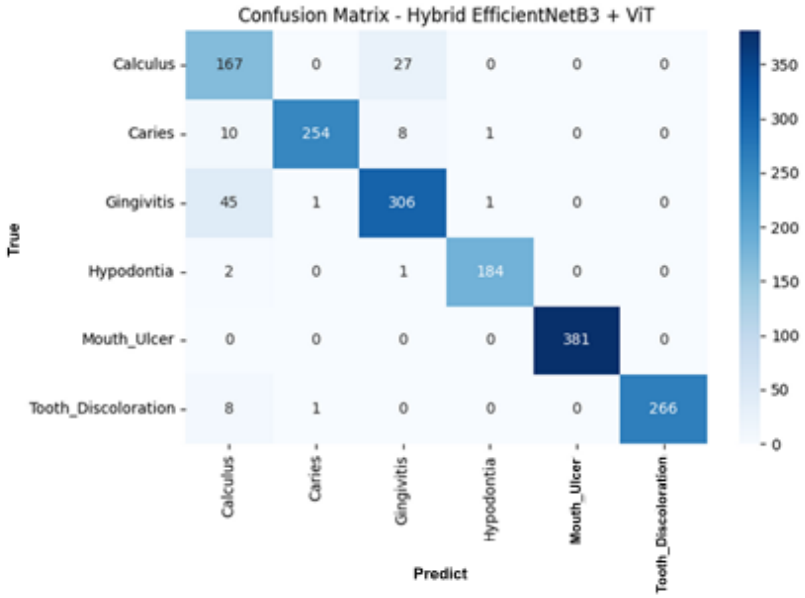
Secara per kelas, Vision Transformer menunjukkan performa yang sangat baik pada kelas *Mouth Ulcer* dengan F1-score sebesar 0,9658, serta pada kelas *Caries* dan *Hypodontia* dengan nilai F1-score masing-masing sebesar 0,9268 dan 0,9222. Namun, performa model relatif lebih rendah pada kelas *Calculus* dan *Gingivitis*, dengan F1-score masing-masing sebesar 0,6174 dan 0,7761, yang mengindikasikan keterbatasan Vision Transformer dalam menangkap detail tekstur lokal tanpa dukungan ekstraksi fitur konvolusional.

Tabel 5. Classification Report Model Vision Transformer

Kelas	Precision	Recall	F1-Score	Support
Calculus	0.5338	0.7320	0.6174	194
Caries	0.9500	0.9048	0.9268	273
Gingivitis	0.8202	0.7365	0.7761	353
Hypodontia	0.9595	0.8877	0.9222	187
Mouth Ulcers	0.9683	0.9633	0.9658	381
Tooth Discolouration	0.9142	0.8909	0.9024	275
Accuracy			0.8581	1663
Macro AVG	0.8577	0.8525	0.8518	1663
Weighted AVG	0.8732	0.8581	0.8634	1663

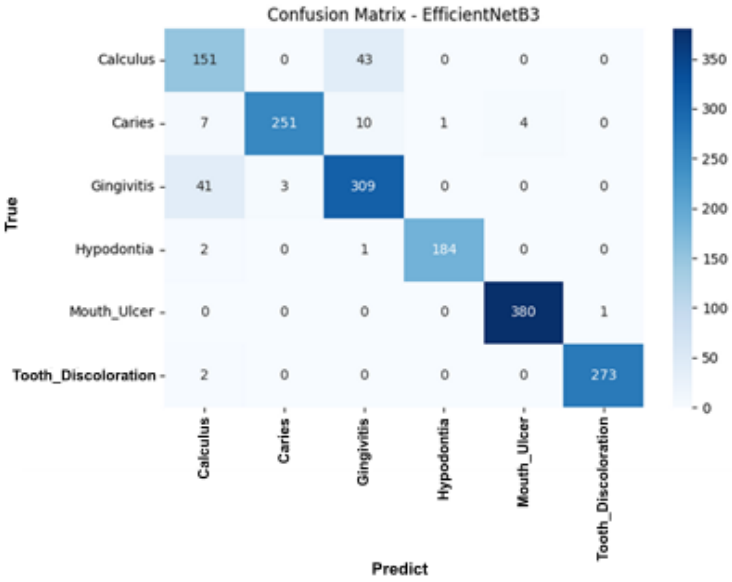
4.2 Analisis *Confusion Matrix*

Berdasarkan analisis *confusion matrix*, sebagian besar citra pada setiap kelas berhasil diklasifikasikan dengan benar oleh *model hybrid*. Kesalahan klasifikasi paling dominan terjadi pada kelas *Calculus*, yang beberapa sampelnya salah diprediksi sebagai *Gingivitis*. Pola kesalahan ini menunjukkan adanya kemiripan karakteristik visual, khususnya pada area plak dan peradangan gusi, yang dapat menimbulkan ambiguitas bagi model. Sebaliknya, kelas *Mouth Ulcer* dan *Hypodontia* menunjukkan tingkat kesalahan yang sangat rendah, yang mengindikasikan bahwa representasi fitur yang dihasilkan oleh *model hybrid* mampu membedakan karakteristik visual kedua kelas tersebut secara jelas. Hasil ini memperkuat bahwa kombinasi fitur lokal dari *EfficientNet-B3* dan pemodelan konteks global oleh *Vision Transformer* berkontribusi positif terhadap peningkatan performa klasifikasi (lihat Gambar 2).



Gambar 2. *Confusion Matrix* Model *Hybrid*

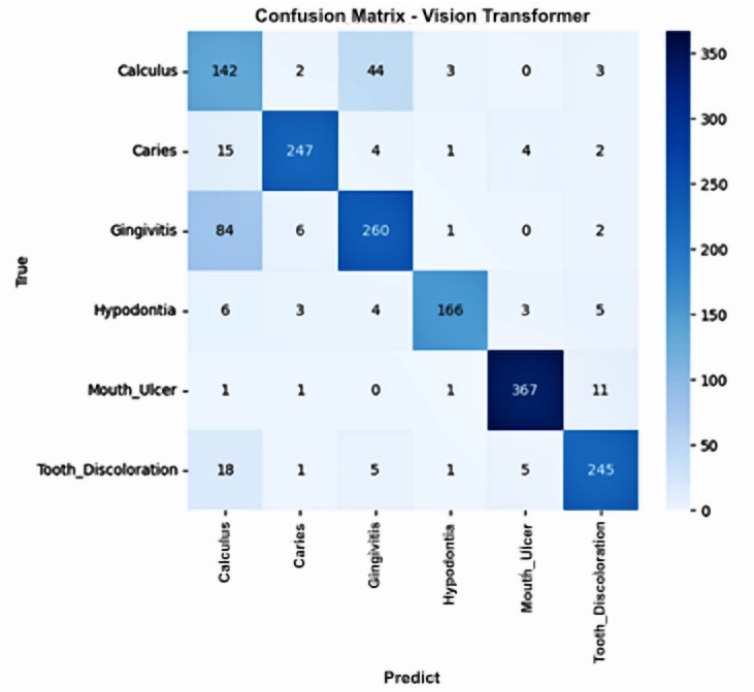
Berdasarkan analisis *confusion matrix*, sebagian besar citra pada setiap kelas berhasil diklasifikasikan dengan benar. Kesalahan klasifikasi yang paling dominan terjadi pada kelas *Calculus*, di mana beberapa sampel salah diprediksi sebagai *Gingivitis*, yang menunjukkan adanya kemiripan visual antara plak gigi dan peradangan gusi sehingga menyulitkan pemisahan kelas secara tegas. Sebaliknya, kesalahan prediksi pada kelas *Mouth Ulcer*, *Hypodontia*, dan *Tooth Discoloration* relatif sangat rendah, yang mengindikasikan bahwa model mampu mengekstraksi fitur tekstur dan struktur visual secara efektif, khususnya melalui kemampuan EfficientNet-B3 dalam menangkap karakteristik lokal yang diskriminatif (lihat Gambar 3).



Gambar 3. *Confusion Matrix* Model EfficientNet

Berdasarkan analisis *confusion matrix*, kesalahan klasifikasi paling signifikan terjadi pada kelas *Gingivitis* yang cukup sering salah diprediksi sebagai *Calculus*, serta sebaliknya pada beberapa sampel *Calculus* yang diprediksi sebagai *Gingivitis*. Pola kesalahan ini mengindikasikan adanya kemiripan visual antara kedua kelas tersebut, khususnya pada area plak dan inflamasi gusi. Meskipun demikian, kelas *Mouth Ulcer* menunjukkan tingkat kesalahan yang sangat rendah, yang menandakan bahwa Vision Transformer mampu mengenali pola lesi yang kontras dan dominan

secara visual. Sementara itu, kesalahan pada kelas *Tooth Discoloration* dan *Hypodontia* relatif terbatas dan tersebar merata tanpa menunjukkan pola kesalahan yang signifikan (lihat Gambar 4).



Gambar 4. *Confusion Matrix* Model Vision Transformer

4.3 Analisis *Wilcoxon Signed-Rank*

Analisis statistik digunakan untuk menegaskan pemilihan model terbaik dalam penelitian ini. Metode analisis statistik yang digunakan adalah *Wilcoxon Signed-Rank* di mana metode ini sudah umum digunakan sebagai alternatif uji nonparametrik terbaik untuk membandingkan performa model klasifikasi [12], [13].

Hasil pengujian statistik dari *Wilcoxon* adalah min(16,5) = 5. Nilai kritis dari n = 6 adalah 2 dan pengujian mendapat hasil W = 5 > 2 dan *p-value* = 0,4375. Hasil W lebih besar daripada 2 dan *p-value* lebih besar dari 0,05, maka kesimpulan yang didapat adalah secara statistic tidak terdapat perbedaan performa yang signifikan antara *Hybrid Model* dan *EfficientNet* (lihat Tabel 6).

Tabel 6. Tabel Wilcoxon Hybrid Model VS EfficientNet

Kelas	Selisih	Rank	Tanda
Calculus	+0.0233	6	+
Gingivitis	+0.0175	5	+
Tooth Discolouration	−0.0111	4	-
Caries	+0.0077	3	+
Mouth Ulcers	+0.0065	2	+
Hypodontia	−0.0026	1	-

Hasil statistik dari *Wilcoxon* adalah min(21,0) = 0. Nilai kritis dari n = 6 adalah 2 dan pengujian mendapat hasil W = 0 ≤ 2 dan *p-value* = 0.03125. Hasil W lebih kecil daripada 2 dan *p-value* lebih kecil dari 0,05, maka kesimpulan yang didapat adalah terdapat perbedaan performa yang signifikan secara statistik antara *Hybrid Model* dan *Vision Transformer*, dengan *Hybrid Model* unggul pada seluruh kelas penyakit (lihat Tabel 7).

Tabel 7. Tabel Wilcoxon Model *Hybrid* VS Vision Transfomer

Kelas	Selisih	Rank	Tanda
Calculus	0.1666	6	+
Gingivitis	0.1045	5	+
Tooth Discolouration	0.0810	4	+
Caries	0.0335	1	+
Mouth Ulcers	0.0342	2	+
Hypodontia	0.0644	3	+

4.4 *Best Selected Model*

Berdasarkan hasil evaluasi dan analisis statistik pada Tabel 6 dan 7, dapat disimpulkan bahwa model *hybrid* EfficientNet–Vision Transformer merupakan model terbaik dalam penelitian ini. Meskipun EfficientNet-B3 menunjukkan capaian numerik tinggi pada beberapa metrik dan Vision Transformer unggul dalam pemodelan konteks global, model *hybrid* mampu memberikan performa yang lebih stabil dan seimbang antar kelas. Model ini mencapai akurasi 93,69% dengan weighted F1-score 93,87% dan macro F1-score 93,25%, yang menunjukkan kemampuan generalisasi yang baik pada *dataset* dengan distribusi kelas tidak seimbang.

Keunggulan model hybrid EfficientNet-Vision Transformer juga sesuai penelitian terdahulu dalam kasus deteksi tumor payudara [14] dan deteksi kanker kolon [15]. Selain itu, kombinasi model dengan metode *CutMix Augmentation* juga sudah teruji dalam penelitian terdahulu untuk kasus klasifikasi kelainan tulang belakang [16] dan penyakit mata [17]. Uji Wilcoxon Signed-Rank menunjukkan bahwa performa model *hybrid* tidak berbeda signifikan secara statistik dengan EfficientNet-B3, tetapi signifikan lebih unggul dibandingkan Vision Transformer.

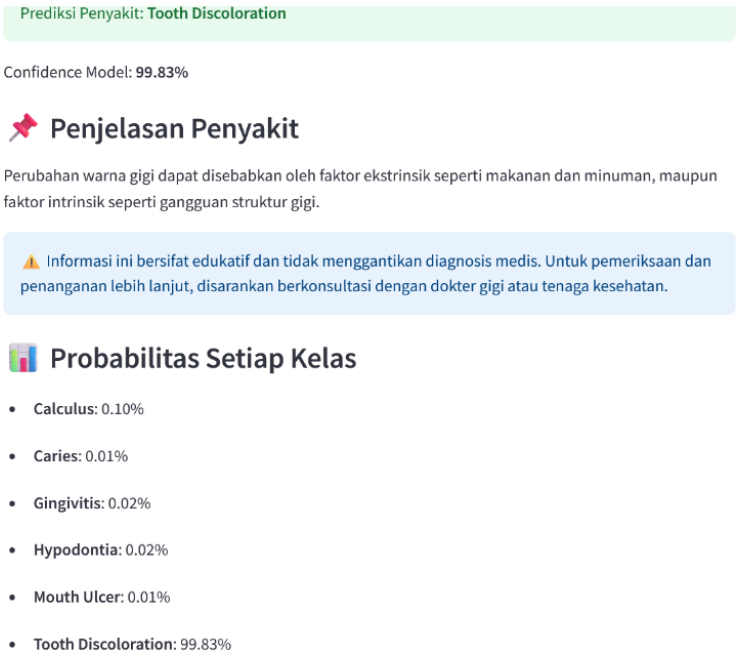
Keunggulan model *hybrid* diperoleh dari kombinasi ekstraksi fitur lokal EfficientNet dan pemodelan konteks global Vision Transformer, yang menghasilkan representasi fitur lebih komprehensif. Temuan ini selaras dengan penelitian terdahulu dan menegaskan efektivitas pendekatan *hybrid* untuk klasifikasi citra penyakit gigi dan mulut.

4.5 Implementasi Model

Implementasi model dilakukan untuk menerapkan *best selected model* ke dalam sebuah sistem yang dapat digunakan secara langsung oleh pengguna. Sistem ini dirancang sebagai aplikasi berbasis web menggunakan framework Streamlit dengan bahasa pemrograman Python, sehingga memungkinkan proses inferensi dilakukan secara interaktif dan *real-time*. Pengguna melakukan input berupa citra penyakit gigi dan mulut melalui antarmuka aplikasi (lihat Gambar 5). Citra yang diunggah kemudian diproses oleh sistem melalui tahapan *preprocessing* yang konsisten dengan proses pelatihan model. Selanjutnya, citra tersebut dimasukkan ke dalam model *hybrid* EfficientNet–Vision Transformer yang telah dipilih sebagai model terbaik.

Gambar 5. Tampilan Antarmuka Sistem

Model menghasilkan keluaran berupa prediksi kelas penyakit gigi dan mulut beserta nilai *confidence score* yang merepresentasikan tingkat keyakinan model terhadap hasil klasifikasi. Hasil prediksi ini ditampilkan secara langsung pada antarmuka aplikasi, sehingga sistem dapat berfungsi sebagai prototipe sistem pendukung diagnosis berbasis citra (lihat Gambar 6). Implementasi ini bertujuan untuk menunjukkan kelayakan model dalam skenario penggunaan nyata serta mempermudah proses evaluasi dan demonstrasi hasil penelitian.



Gambar 6. Tampilan *Output* Sistem

5. Kesimpulan dan Saran

Penelitian ini berhasil mengembangkan model *hybrid* EfficientNet–Vision Transformer yang mengintegrasikan ekstraksi fitur lokal dan konteks global dengan performa tinggi, mencapai

akurasi sebesar 93,69%. Berdasarkan analisis statistik menggunakan uji Wilcoxon Signed-Rank, model *hybrid* terbukti unggul secara signifikan dibandingkan Vision Transformer tunggal (*p-value* < 0,05) dan menunjukkan konsistensi performa yang lebih stabil daripada EfficientNet-B3, terutama pada kelas kompleks seperti Calculus dan Gingivitis. Implementasi pada aplikasi Streamlit juga mengonfirmasi bahwa model ini valid secara eksperimental dan fungsional untuk klasifikasi *real-time*, menjadikannya prototipe sistem pendukung Keputusan yang layak. Untuk penelitian lebih lanjut, disarankan untuk memperluas keragaman dan keseimbangan *dataset* guna meningkatkan generalisasi serta meminimalkan bias prediksi. Dari sisi teknis, eksplorasi variasi arsitektur Transformer yang lebih kompleks dan teknik penanganan data tidak seimbang perlu diperdalam. Selain itu, pengembangan sistem di masa depan sebaiknya menyertakan fitur validasi kualitas citra *input* dan integrasi lebih lanjut dengan tenaga medis untuk memastikan keamanan serta relevansi klinis sebelum diterapkan pada skala yang lebih luas.

Referensi

[1] Kementerian Kesehatan Republik Indonesia, "Factsheet Kesehatan Gigi dan Mulut (Gilut)," 2023, *Badan Kebijakan Pembangunan Kesehatan*. [Online]. Available: https://repository.badankebijakan.kemkes.go.id/id/eprint/5534/1/04 factsheet Gilut_bahasa.pdf

[2] A. P. Sinaga, A. R. Ismail, M. A. P. Siregar, and I. D. Saraswati, "Korelasi Disparitas Ketersediaan Tenaga Medis Gigi Antardaerah Terhadap Pemanfaatan Layanan Gigi dan Mulut di Indonesia," *J. Manaj. Pelayanan Kesehat.*, vol. 25, no. 4, pp. 217–224, 2022, [Online]. Available: https://www.researchgate.net/publication/366535008_Korelasi_Disparitas_Ketersediaan_Tenaga_Medis_Gigi_Antardaerah_Terhadap_Pemanfaatan_Layanan_Gigi_dan_Mulut_di_Indonesia

[3] M. Harahap and A. M. Husein, "Peneraparan Efficient-Net dalam mengklasifikasi kanker kulit," *Pros. Semin. Nas. Ilmu Komput.*, Jul. 2024, [Online]. Available: https://jurnal.unprimdn.ac.id/index.php/isbn/article/view/5405

[4] J. Yang *et al.*, "Focal self-attention for Local-Global Interactions in Vision Transformers," *arXiv Prepr. arXiv2107.00641*, Jul. 2021, [Online]. Available: https://arxiv.org/abs/2107.00641

[5] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features," *arXiv Prepr. arXiv1905.04899*, May 2019, [Online]. Available: https://arxiv.org/abs/1905.04899

[6] W. Wahyuningsih, G. S. Nugraha, and R. Dwiyansaputra, "Classification Of Dental Caries Disease In Tooth Images Using A Comparison Of Efficientnet-B0, Mobilenetv2, Resnet-50, Inceptionv3 Architectures," *J. Tek. Inform.*, vol. 5, no. 4, pp. 177–185, Jul. 2024.

[7] F. Lavenia, C. M. S. Ramdani, and I. Hoeronis, "Klasifikasi Penyakit Pulpitis Pada Citra Radiografi Periapikal Menggunakan Metode Convolutional Neural Network (CNN)," *Media J. Inform.*, vol. 16, no. 1, Jun. 2024, doi: 10.35194/mji.v16i1.4098.

[8] A. N. Pratama, "Sistem Klasifikasi Penyakit Kulit pada Manusia Convolutional Neural Network (CNN) EfficientNet B2," *Paradig. - J. Komput. dan Inform.*, vol. 30, no. 2, Jun. 2024, doi: 10.33503/paradigma.v30i2.439.

[9] R. R. Ar, "pendeteksian dini stunting pada balita menggunakan vision transformer (vit) berbasis citra tubuh," *J. Inform. dan Tek. Elektro Terap.*, vol. 13, no. 3S1, 2025, doi: 10.23960/jitet.v13i3s1.7888.

[10] I. Yudistiansyah, "Penerapan Computer Vision Untuk Klasifikasi Penyakit Mata Menggunakan Arsitektur Vision Transformers (Vits) Pada Citra Fundus," Nusa Putra University, 2025.

[11] S. Sajid, "Oral Disease Dataset." [Online]. Available: https://www.kaggle.com/datasets/salmansajid05/oral-diseases/data

[12] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep.*, vol. 14, no. 1, pp. 1–14, 2024, doi: 10.1038/s41598-024-56706-x.

[13] G. Airlangga, "Predicting Diabetes with Machine Learning: Evaluating Tree-Based and Ensemble Models with Custom Metrics and Statistical Validation," *Build. Informatics, Technol. Sci.*, vol. 6, no. 3, pp. 1818–1827, 2024, doi: 10.47065/bits.v6i3.6419.

[14] A. Singh, S. P. Mishra, P. Singh, and A. Srivastava, "VISNET: An Efficient Light Weighted Hybrid Model for Early Detection of Breast Tumour in Ultrasound Images using Vision Transformer and Convolutional Neural Networks," *J. Inf. Syst. Eng. Manag.*, 2025, doi: 10.52783/jisem.v10i40s.9215.

[15] B. Sathyanarayana, S. Alampally, R. Akella, and V. V. R. Indugu, "ColoViT: a synergistic integration of EfficientNet and vision transformers for advanced colon cancer detection," *J. Cancer Res. Clin. Oncol.*, vol. 151, no. 7, pp. 1–19, 2025, doi: 10.1007/s00432-025-06199-6.

[16] J. F. M. Pereira, J. F. Mari, and L. H. F. P. Silva, "Exploiting Data Augmentation Strategies to Improve the Classification of Spinal Disorders in X-Ray Images," *Rev. Inform. Teor. e Apl.*, vol. 32, no. 1, pp. 257–264, 2025, doi: 10.22456/2175-2745.143521.

[17] X. Qi *et al.*, "MediAug: Exploring Visual Augmentation in Medical Imaging," *Lect. Notes Comput. Sci.*, vol. 15916 LNCS, pp. 218–232, 2026, doi: 10.1007/978-3-031-98688-8_16.