Image Captioning untuk Diagnosis Medis Berbasis Citra Serviks

Ridwan Akrom
Jurusan Sistem Komputer
Fakultas Ilmu Komputer
Universitas Sriwijaya
Palembang, Indonesia

Intelligent System Research Group
Jurusan Sistem Komputer
Fakultas Ilmu Komputer
Universitas Sriwijaya
Palembang, Indonesia
ridwanakrom020204@gmail.com

Muhammad Naufal Rachmatullah*
Intelligent System Research Group
Jurusan Sistem Komputer
Fakultas Ilmu Komputer
Universitas Sriwijaya
Palembang, Indonesia
naufalrachmatullah@gmail.com

Akhiar Wista Arum
Intelligent System Research Group
Jurusan Sistem Komputer
Fakultas Ilmu Komputer
Universitas Sriwijaya
Palembang, Indonesia
wistaarum16@gmail.com

Anggun Islami
Intelligent System Research Group
Jurusan Sistem Komputer
Fakultas Ilmu Komputer
Universitas Sriwijaya
Palembang, Indonesia
anggunislami2@gmail.com

Ade Iriani Sapitri
Intelligent System Research Group
Jurusan Sistem Komputer
Fakultas Ilmu Komputer
Universitas Sriwijaya
Palembang, Indonesia
adeirianisapitri13@gmail.com

***Ringkasan*—Deteksi dini kanker serviks sangat penting untuk mencegah perkembangan kanker serviks pada pasien, namun skrining manual saat ini memiliki keterbatasan signifikan terkait subjektivitas, waktu, dan biaya. Penelitian ini bertujuan untuk menerapkan *image captioning* sebagai alat bantu diagnosis berbasis *deep learning* yang mampu menghasilkan laporan medis deskriptif dari data citra serviks. Metode yang diusulkan menggunakan arsitektur *encoder-decoder* yang mengombinasikan *EfficientNetV2M* untuk ekstraksi fitur visual dan arsitektur *Transformer* sebagai *encoder-decoder* untuk generasi teks. *EfficientNetV2M* dipilih karena efisiensinya dalam ekstraksi fitur, sementara *Transformer* digunakan karena kemampuannya yang unggul dalam menangkap dependensi kontekstual melalui mekanisme *self-attention*. Model dilatih dan dievaluasi menggunakan data citra serviks yang telah memiliki laporan medis sebagai *ground truth*. Hasil evaluasi kuantitatif menunjukkan performa yang menjanjikan, dengan model berhasil mencapai skor yang kompetitif pada metrik *Bilingual Evaluation Understudy* (BLEU) dan *Metric for Evaluation of Translation with Explicit Ordering* (METEOR), yaitu 0.677 (BLEU-1), 0.6038 (BLEU-2), 0.5570 (BLEU-3), dan 0.5189 (BLEU-4), serta 0.7038 pada METEOR. Selain itu, analisis kualitatif menunjukkan kemampuan model dalam mengidentifikasi atribut-atribut klinis utama dengan benar, meskipun masih ditemukan ketidakakuratan minor pada detail visual yang sangat halus. Penelitian ini menunjukkan bahwa arsitektur *EfficientNet-Transformer* merupakan pendekatan yang efektif dan berpotensi untuk dikembangkan sebagai alat bantu diagnosis, guna meningkatkan konsistensi dan efisiensi dalam pelaporan medis kanker serviks.**

***Keywords*— *Deep Learning*, *EfficientNet*, *Image Captioning*, Kanker Serviks, Laporan Medis Otomatis, *Transformer*.**

I. PENDAHULUAN

Kanker serviks adalah jenis kanker yang berkembang pada leher rahim dan saat ini menempati peringkat keempat sebagai penyebab kematian paling umum pada wanita di dunia [1][2]. Banyak cara yang dilakukan untuk mendiagnosis penyakit ini, salah satunya melalui skrining seperti tes HPV, *pap smear* maupun inspeksi visual dengan asam asetat (IVA) [3]. Namun, proses diagnosis secara manual melalui interpretasi citra medis sangat bergantung pada kemampuan, pengalaman, serta tingkat konsentrasi pemeriksa. Kondisi ini dapat meningkatkan kemungkinan terjadinya kesalahan diagnosis, terutama pada tahap awal ketika perubahan sel masih sulit dibedakan [4]. Untuk mengatasi keterbatasan tersebut, teknologi *Artificial Intelligence* (AI) dengan pendekatan *Deep Learning* mulai banyak dimanfaatkan dalam analisis citra medis hasil skrining. Beberapa tahun terakhir, banyak algoritma *Deep Learning* diusulkan untuk analisis citra medis yang terbukti mampu mengurangi kesalahan diagnosis manual dan meningkatkan akurasi [5]. Salah satunya adalah metode *Image Captioning*, sebuah solusi inovatif untuk menghasilkan deskripsi tekstual yang informatif dari citra secara otomatis. Seperti yang dilakukan oleh [6][7], penelitian tersebut menerapkan *Image Captioning* berbasis model *Vision Image Transformer* (ViT) sebagai *encoder* dan *Generative Pretrained Transformer* (GPT) sebagai *decoder* pada data medis. Pendekatan ini menunjukkan bahwa

integrasi antara model visi dan bahasa mampu menghasilkan deskripsi medis yang lebih akurat dan kontekstual.

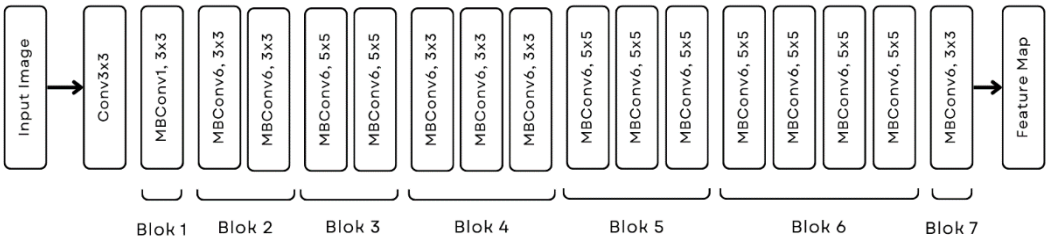
Selanjutnya, penelitian dalam [8] memperluas pendekatan ini dengan memanfaatkan model *EfficientNet-Transformer* pada tugas pembuatan laporan medis dari citra rontgen dada. Hasilnya menunjukkan peningkatan skor metrik evaluasi yang signifikan sekaligus ketahanan terhadap *overfitting*, menandakan potensi besar arsitektur ini dalam ekstraksi fitur visual medis.

Meskipun demikian, hingga saat ini belum banyak penelitian yang secara spesifik menerapkan *EfficientNet* dalam tugas *Image Captioning* pada domain citra medis, khususnya pada citra serviks. Hal ini menjadi celah penelitian yang menarik untuk dieksplorasi lebih lanjut. Oleh karena itu, penelitian ini berfokus pada penerapan *EfficientNetV2M* yang dikombinasikan dengan *Transformer* untuk melakukan generasi diagnosis citra serviks guna pendeteksian dini dalam mencegah kanker serviks. Dalam penelitian ini, untuk mengukur performa model dalam generasi diagnosis diukur dengan menggunakan metrik evaluasi BLEU dan METEOR. Hasil penelitian ini direncanakan mampu berkontribusi mengurangi kesalahan diagnosis secara manual dan meringankan beban kerja radiolog, serta memberikan manfaat untuk penelitian selanjutnya.

II. METODOLOGI

A. *EfficientNet*

EfficientNet merupakan arsitektur dari CNN yang modern dan sangat efisien ketika digunakan untuk tugas klasifikasi citra maupun ekstraksi fitur [9]. *EfficientNet* menjadi salah satu arsitektur yang populer digunakan sebagai *encoder* untuk mengekstraksi fitur visual dari sebuah citra. Arsitektur ini memiliki satu prinsip utama yang menjadi keunggulannya yaitu *Compound Scaling* [10]. Arsitektur model *EfficientNetV2M* dapat dilihat pada Gambar 1.



Gambar 1 Arsitektur *EfficientNetV2M*

Pada proses *Compound Scaling*, dilakukan penskalaan model secara seragam dan seimbang pada tiga dimensi utama: kedalaman jaringan (*depth*), lebar jaringan (*width*), dan resolusi citra (*resolution*). Selain itu, pada setiap tahapnya, penskalaan ini diatur oleh sebuah koefisien kompon (ϕ) yang memastikan ketiga dimensi tersebut tumbuh secara proporsional. Arsitektur *EfficientNet* memastikan perolehan akurasi yang lebih tinggi dengan efisiensi komputasi yang lebih baik dibandingkan dengan arsitektur yang hanya menskalakan satu dimensi saja [11].

Sedangkan untuk prinsip penskalaannya, hal penting yang dilakukan adalah menggunakan serangkaian koefisien tetap untuk menyeimbangkan hubungan antara ketiga dimensi tersebut. *EfficientNet* menggunakan aturan penskalaan matematis sebagai dasar evaluasi dalam arsitekturnya. Aturan penskalaan tersebut dapat dilihat pada persamaan (1), (2), (3), dan (4).

$$\text{depth: } d = \alpha\phi \quad (1)$$

$$\text{width: } w = \beta\phi \quad (2)$$

$$\text{resolution: } r = \gamma\phi \quad (3)$$

$$\text{s.t. } \alpha \cdot \beta 2 \cdot \gamma 2 \approx 2 \quad (4)$$

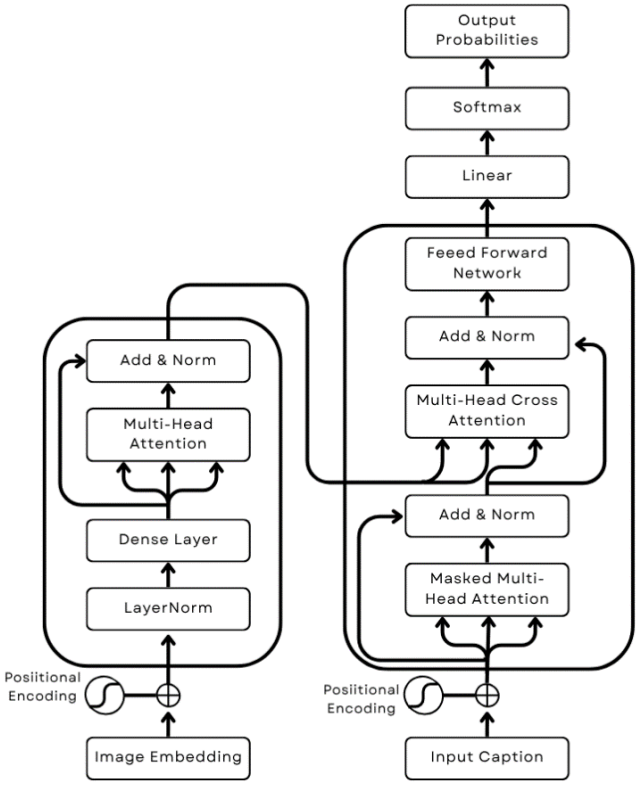
Dari persamaan (1), (2), (3), dan (4), *Compound Scaling* ditentukan oleh tiga variabel utama. Variabel α, β, dan γ merupakan konstanta yang mengatur seberapa besar penskalaan untuk kedalaman, lebar, dan resolusi. Sementara itu, ϕ adalah koefisien yang ditentukan pengguna untuk mengontrol seberapa besar model akan diperbesar dari model dasarnya. *Compound Scaling* pada arsitektur *EfficientNet* digunakan untuk menjaga keseimbangan antar dimensi jaringan, yang bertujuan untuk menghasilkan ekstraksi fitur yang lebih optimal dan efisien secara komputasi [12].

B. *Transformer*

Transformer merupakan arsitektur yang sangat populer dalam bidang *Natural Language Processing* (NLP) [13]. *Transformer* adalah salah satu arsitektur yang dominan digunakan untuk melakukan tugas *sequence-to-sequence* seperti generasi teks [14]. *Transformer* mempunyai dua poin penting dalam arsitekturnya yaitu struktur *encoder-decoder* dan mekanisme *Self-Attention* [15]. Gambar 2 menyajikan diagram arsitektur *Transformer* secara umum yang digunakan dalam penelitian ini.

Mekanisme *attention* dihitung menggunakan persamaan (5). Dengan berdasarkan tiga representasi dari data masukan. Di mana Q adalah matriks *Query*, K adalah matriks *Key*, dan V adalah matriks *Value*, sementara $\sqrt{d_k}$ adalah faktor penskalaan. *Attention* pada arsitektur *Transformer* digunakan untuk menghitung skor relevansi antara setiap elemen dalam urutan data, sehingga model dapat secara dinamis menentukan bagian mana dari masukan yang paling penting untuk diperhatikan saat menghasilkan keluaran [16].

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_K}}\right)V \quad (5)$$



Gambar 2 Arsitektur *Transformer*

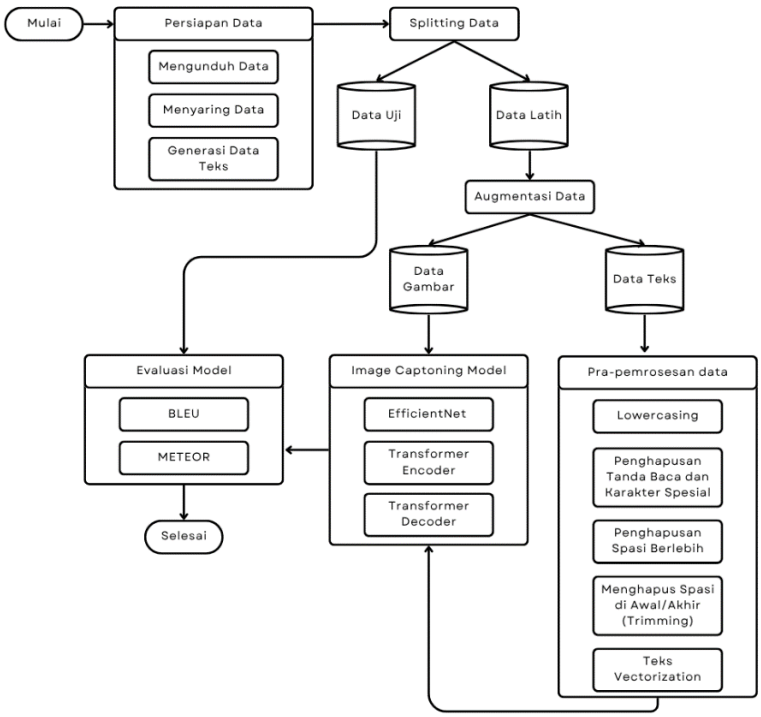
III. EKSPERIMEN DAN HASIL

Bagian ini menguraikan secara rinci alur eksperimen yang dilakukan dalam penelitian. Proses diawali dengan tahap persiapan data, yang mencakup pengunduhan, penyaringan, dan proses generasi teks menjadi kalimat untuk dijadikan label dari data gambar. Lalu, pembagian dataset menjadi data latih dan data uji. Berikutnya, data latih diperkaya melalui augmentasi dan dilanjutkan data teks melalui serangkaian langkah pra-pemrosesan, Model *image captioning* yang diusulkan, dengan arsitektur berbasis *EfficientNet* dan *Transformer*, kemudian dilatih. Pada tahap akhir, kinerja model dievaluasi secara kuantitatif terhadap data uji menggunakan metrik BLEU dan METEOR untuk mengukur kualitas teks yang dihasilkan. Keseluruhan tahap ini dirangkum secara visual pada Gambar 3.

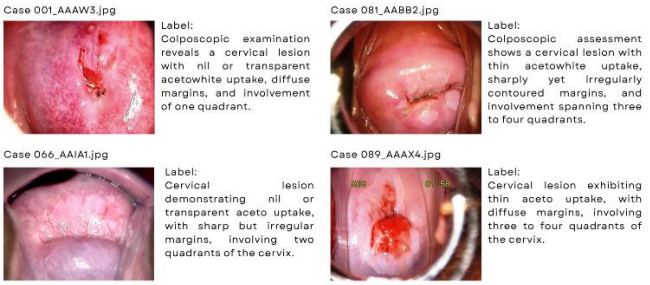
A. Dataset dan Persiapan data

Dalam penelitian ini, kami berfokus pada penggunaan data citra serviks hasil skrining kolposkopi dari *International Agency for Research on Cancer* (IARC) yang teridentifikasi *after acetic acid* [17]. Dataset ini mencakup pasangan gambar kolposkopi dan metadata teks deskriptif dari 200 pasien.

Berdasarkan Gambar 3, penelitian dimulai dengan proses pengunduhan data melalui situs web resmi IARC. Setelah data dikumpulkan, berikutnya data tersebut disaring untuk mendapatkan data gambar *after acetic acid.* Setelah data disaring, diperoleh pasangan data gambar dan teks sebanyak 297 pasang. Pada metadata, dilakukan pemilahan fitur yang relevan yakni *Aceto uptake, Margins*, dan *Lesion size.* Lalu, fitur tersebut digabungkan dan digenerasi agar menjadi satu kalimat baku sebagai label dari data gambar. Gambar 4 merupakan sampel dari data yang digunakan.



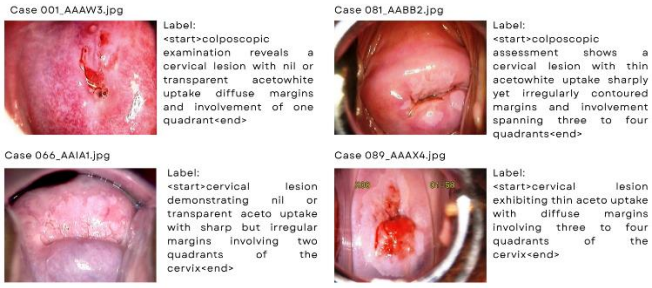
Gambar 3 Alur kerja yang digunakan dalam penelitian



Gambar 4 Sampel data serviks kolposkopi IARC

Pra-Pemrosesan

Pada tahap ini, data yang telah disiapkan akan diproses lebih lanjut agar siap digunakan sebagai masukan untuk model yang diusulkan. Dalam tahap ini data dibagi berdasarkan *Case* terlebih dahulu menjadi data latih dan data uji dengan perbandingan 80:20. Sehingga diperoleh data latih sebanyak 235 pasang dan data uji sebanyak 62 pasang. Lalu, augmentasi diterapkan pada data latih dengan merotasi hingga 15 derajat, *horizontal and vertical flip*. Augmentasi perlu dilakukan karena data yang ada hanya sedikit, sehingga perlu dilakukan augmentasi untuk memperkuat pembelajaran model agar menghasilkan model yang kokoh [18]. Berikutnya, pemrosesan data teks dan gambar yang meliputi, *lowercasing*, *punctuations removal*, menambah token <*start*> & <*end*>, vektorisasi teks, *decode* dan *resize* gambar ke ukuran 224x224. Keseluruhan tahap pra-pemrosesan ini dilakukan dengan *library* standar python. Gambar 5 menyajikan contoh data yang telah melalui tahap pra-pemrosesan.



Gambar 5 Sampel data hasil pra-pemrosesan

B. Training dan Testing

Proses pelatihan model memanfaatkan dataset yang telah melalui tahap pra-pemrosesan. Data ini dibagi dengan perbandingan 80:20, menghasilkan 235 pasangan citra-teks sebagai data latih dan 62 pasangan sebagai data uji. Model dikonfigurasi untuk memproses citra masukan dengan ukuran 224x224 piksel dan bekerja dengan ukuran *vocab size* sebanyak 10.000 token, serta panjang sekuens maksimum 27 token. Proses pelatihan dijalankan selama 150 *epoch* menggunakan *optimizer Adam* dengan *learning rate* 1e-5. Pelatihan dilakukan dengan ukuran *batch* 64 dan menerapkan *callback EarlyStopping* untuk mencegah *overfitting* serta menghentikan pelatihan pada *epoch* terbaik.

C. Metrik Evaluasi dan Hasil

Proses evaluasi dilakukan dengan membandingkan teks yang dihasilkan oleh model terhadap teks referensi (*ground truth*) dari data uji. Kinerja model diukur secara kuantitatif menggunakan metrik evaluasi standar, yakni BLEU dan METEOR. Rincian lengkap mengenai kedua metrik tersebut dijelaskan pada bagian dibawah ini.

BLEU

BLEU adalah metrik evaluasi otomatis yang digunakan untuk mengukur kesamaan antara teks kandidat dengan teks referensi [19][20][21]. BLEU memberikan indikasi kualitas dengan mengukur kecukupan makna (*adequacy*) dan kelancaran bahasa (*fluency*) melalui presisi n-gram. Namun, metrik ini bisa menyesatkan karena hanya menggunakan presisi tanpa memperhitungkan *recall*, tidak dapat menghitung kata dengan akar kata yang sama atau sinonim, dan sering kali gagal merefleksikan makna serta struktur kalimat yang sesungguhnya [22]. Metrik ini dihitung menggunakan persamaan (6).

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} \omega log\ P_n) \quad (6)$$

METEOR

METEOR adalah metrik yang didasarkan pada konsep umum pencocokan unigram antara teks kandidat dan referensi [23]. Pencocokan ini dilakukan pada tiga level, yaitu pencocokan bentuk permukaan kata atau unigram kata secara persis, pencocokan bentuk dasar kata atau stem, dan pencocokan makna kata melalui sinonim [24][25]. Skor METEOR kemudian dihitung menggunakan kombinasi dari presisi dan *recall* dari hasil pencocokan unigram tersebut, serta sebuah pengukuran yang melibatkan n-gram terpanjang yang sama-sama ditemukan pada kalimat referensi dan kandidat. Rincian perhitungan metrik ini dapat dilihat pada persamaan (7).

$$METEOR = F_mean \cdot (1 - Penalty) \quad (7)$$

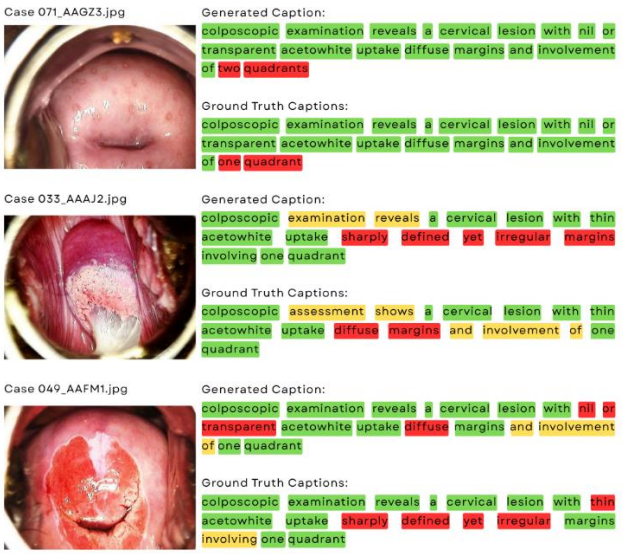
Hasil

Hasil evaluasi kuantitatif dari metode yang diusulkan dirangkum dalam Tabel I. Tabel ini menyajikan skor yang dicapai model pada berbagai metrik evaluasi yang diusulkan yakni BLEU dan METEOR.

Tabel I Hasil evaluasi model

Parameter	Metrik Evaluasi	Skor
Image size = 224x224, *Vocab size* = 10.000, *Sequence length* = 27, *Batch size* = 64, *Epoch* = 150, *Optimizer* = *Adam*, *Callback* = *EarlyStopping*	BLEU-1	0.6770
	BLEU-2	0.6038
	BLEU-3	0.5570
	BLEU-4	0.5189
	METEOR	0.7038

Selain evaluasi kuantitatif pada Tabel II, ilustrasi performa kualitatif dari model yang diusulkan dapat dilihat pada Gambar 6. Gambar ini menyajikan perbandingan visual antara teks yang dihasilkan oleh model dengan teks referensi (*ground truth*).



Gambar 6 Contoh hasil generasi model

IV. DISKUSI

Hasil penelitian yang diringkas dalam Tabel II menunjukkan skor BLEU dan METEOR yang kompetitif, membuktikan efektivitas arsitektur yang diusulkan untuk domain medis. Analisis kualitatif mendalam juga dilakukan untuk mengevaluasi luaran teks dari arsitektur *EfficientNet-Transformer*, yang diilustrasikan secara visual pada Gambar 6. Dalam analisis ini, diterapkan sebuah skema kode warna untuk memudahkan perbandingan dengan *ground truth*. Teks yang ditandai hijau menunjukkan kesesuaian penuh dan identifikasi atribut yang akurat. Teks kuning merepresentasikan penggunaan frasa yang berbeda namun memiliki makna yang sama atau sinonim, yang menyoroti kemampuan model dalam variasi bahasa. Di sisi lain, teks merah mengindikasikan adanya deskripsi yang salah atau gagal dideteksi oleh model.

Untuk memberikan konteks performa yang lebih luas, Tabel II menyajikan komparasi metrik evaluasi model kami dengan beberapa penelitian terkini di bidang pembuatan laporan citra medis.

Tabel II Perbandingan metrik evaluasi dengan penelitian sebelumnya

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Model-SEY [6]	0.6412	0.5335	0.4395	0.3716	-
VIT-GPT [7]	0.403	0.296	0.178	0.127	0.164
EfficientNetB1 [8]	0.48	0.29	0.19	0.14	-
EfficientNet-Transformer	**0.6770**	**0.6038**	**0.5570**	**0.5189**	**0.7038**

Tabel II menunjukkan bahwa arsitektur *EfficientNet-Transformer* kami secara konsisten mengungguli model pembanding. Model kami mencapai skor BLEU-4 0.5189 dan METEOR 0.7038, secara signifikan melampaui model terbaik kedua, Model-SEY [6] (BLEU-4 0.3716), dan VIT-GPT [7] (METEOR 0.164). Pencapaian ini membuktikan bahwa penggunaan arsitektur *EfficientNet-Transformer* berkontribusi signifikan terhadap peningkatan kualitas *Image Captioning*, terutama dalam domain medis. Secara keseluruhan, hasil analisis memperlihatkan bahwa model berhasil menangkap konteks esensial dan elemen-elemen penting citra, sekalipun masih ditemukan sedikit ketidakakuratan pada detail klinis yang sangat halus.

V. KESIMPULAN

Penelitian ini telah berhasil mengembangkan sebuah model *Image Captioning* untuk citra serviks guna menghasilkan laporan medis otomatis. Model ini didasarkan pada arsitektur *encoder-decoder* yang efisien, dengan mengimplementasikan *EfficientNet* untuk ekstraksi fitur visual dan *Transformer* sebagai *encoder-decoder* untuk generasi teks.

Berdasarkan hasil evaluasi, arsitektur yang diusulkan menunjukkan performa yang sangat menjanjikan. Secara kuantitatif, model ini berhasil mencapai skor yang tinggi pada berbagai metrik evaluasi standar, termasuk BLEU dan METEOR. Capaian ini mengindikasikan bahwa deskripsi yang dihasilkan oleh model memiliki kesamaan n-gram dan keselarasan semantik yang baik dengan teks *ground truth.*

Secara kualitatif, analisis visual menunjukkan bahwa model mampu mengidentifikasi atribut-atribut klinis utama yang relevan dari citra serviks, seperti karakteristik *aceto uptake* dan *lesion size*. Meskipun demikian, penelitian ini juga mencatat adanya ketidakakuratan minor dalam mendeskripsikan detail visual yang sangat halus dan spesifik, seperti penentuan *margins*. Hal ini menunjukkan bahwa model telah berhasil menangkap konteks utama citra dengan baik, namun masih memerlukan peningkatan dalam membedakan nuansa klinis yang subtil.

Secara keseluruhan, penelitian ini membuktikan bahwa kombinasi *EfficientNet-Transformer* adalah pendekatan yang efektif dan berpotensi besar untuk dikembangkan sebagai sistem alat bantu untuk mendukung dan meringankan beban kerja tenaga medis dalam pembuatan laporan. Untuk

penelitian selanjutnya, disarankan untuk melakukan pelatihan dengan dataset yang lebih besar dan beragam, serta mengeksplorasi teknik *fine-tuning* yang lebih lanjut guna meningkatkan akurasi model pada detail-detail klinis yang spesifik.

REFERENSI

[1] S. Nurmaini *et al.*, "Real time mobile AI-assisted cervicography interpretation system," *Informatics Med. Unlocked*, vol. 42, no. September, p. 101360, 2023, doi: 10.1016/j.imu.2023.101360.

[2] N. Youneszade, M. Marjani, and C. P. Pei, "Deep Learning in Cervical Cancer Diagnosis: Architecture, Opportunities, and Open Research Challenges," *IEEE Access*, vol. 11, no. December 2022, pp. 6133–6149, 2023, doi: 10.1109/ACCESS.2023.3235833.

[3] S. J. Davidson, D. El-Rayes, M. Khalifa, C. S. Fok, and B. K. Erickson, "HPV-independent cervical cancer associated with non-reducible pelvic organ prolapse: A case report," *Gynecol. Oncol. Reports*, vol. 53, no. February, p. 101408, 2024, doi: 10.1016/j.gore.2024.101408.

[4] R. Austin and R. Parvathi, "CNN based method for classifying cervical cancer cells in pap smear images," *Sci. Rep.*, vol. 15, no. 1, pp. 1–21, 2025, doi: 10.1038/s41598-025-10009-x.

[5] M. K. Nour, I. Issaoui, A. Edris, A. Mahmud, M. Assiri, and S. S. Ibrahim, "Computer Aided Cervical Cancer Diagnosis Using Gazelle Optimization Algorithm with Deep Learning Model," *IEEE Access*, vol. 12, no. December 2023, pp. 13046–13054, 2024, doi: 10.1109/ACCESS.2024.3351883.

[6] M. Ucan, B. Kaya, and M. Kaya, "Turkish Chest X-Ray Report Generation Model Using the Swin Enhanced Yield Transformer (Model-SEY) Framework," *Diagnostics*, vol. 15, no. 14, pp. 1–13, 2025, doi: 10.3390/diagnostics15141805.

[7] H. Tsaniya, C. Fatichah, N. Suciati, T. Obi, and J. S. Lee, "Medical Report Generation with Knowledge Distillation and Multi-Stage Hierarchical Attention in Vision Transformer encoder and GPT-2 decoder," *IEEE Access*, vol. 13, no. July, pp. 132973–132989, 2025, doi: 10.1109/ACCESS.2025.3588344.

[8] M. M. Henry, N. A. Heryanto, and B. Pardamean, "Exploring EfficientNet Variants for Image Encoding in Auxiliary Signal Guided Knowledge Encoder-Decoder Framework," *Procedia Comput. Sci.*, vol. 245, no. C, pp. 391–398, 2024, doi: 10.1016/j.procs.2024.10.265.

[9] P. Malik, "Deep Learning Architectures for Image Recognition : a Deep Learning Architectures for Image Recognition : a Comprehensive," no. September, 2023.

[10] R. Farkh, G. Oudinet, and Y. Foued, "Image Captioning Using Multimodal Deep Learning Approach," *Comput. Mater. Contin.*, vol. 81, no. 3, pp. 3951–3968, 2024, doi: 10.32604/cmc.2024.053245.

[11] R. Raza *et al.*, "Lung-EffNet: Lung cancer classification using EfficientNet from CT-scan images," *Eng. Appl. Artif. Intell.*, vol. 126, no. PB, p. 106902, 2023, doi: 10.1016/j.engappai.2023.106902.

[12] S. J. Colaco and D. S. Han, "Deep Learning-Based Facial Landmarks Localization Using Compound Scaling," *IEEE Access*, vol. 10, pp. 7653–7663, 2022, doi: 10.1109/ACCESS.2022.3141791.

[13] W. Ansar, S. Goswami, and A. Chakrabarti, "A Survey on Transformers in NLP with Focus on Efficiency," pp. 1–31, 2024, [Online]. Available: http://arxiv.org/abs/2406.16893

[14] S. Masih, M. Hassan, L. G. Fahad, and B. Hassan, "Transformer-Based Abstractive Summarization of Legal Texts in Low-Resource Languages," *Electron.*, vol. 14, no. 12, pp. 1–21, 2025, doi: 10.3390/electronics14122320.

[15] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, vol. 3, no. October, pp. 111–132, 2022, doi: 10.1016/j.aiopen.2022.10.001.

[16] M. H. Guo *et al.*, "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, no. 3, pp. 331–368, 2022, doi: 10.1007/s41095-022-0271-y.

[17] Priyadarshini Chatterjee and Shadab Siddiqui, "IARC Image ColpoBank," IEEE Dataport. Accessed: Aug. 08, 2025. [Online]. Available: https://ieee-dataport.org/documents/iarc-image-colpobank#:~:text=The Dataset is aquired from IARC Image Colpo,when results from a Pap test are abnormal.

[18] R. Bravin, L. Nanni, A. Loreggia, S. Brahnam, and M. Paci, "Varied Image Data Augmentation Methods for Building Ensemble," *IEEE Access*, vol. 11, no. December 2022, pp. 8810–8823, 2023, doi: 10.1109/ACCESS.2023.3239816.

[19] C. Dong *et al.*, "A Survey of Natural Language Generation," *ACM Comput. Surv.*, vol. 55, no. 8, pp. 1–36, 2023, doi: 10.1145/3554727.

[20] T. C. Tong, S. He, Z. Shao, and D. Y. Yeung, "G-VEval: A Versatile Metric for Evaluating Image and Video Captions Using GPT-4o," *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 7, pp. 7419–7427, 2025, doi: 10.1609/aaai.v39i7.32798.

[21] T. Pang, P. Li, and L. Zhao, "A survey on automatic generation of medical imaging reports based on deep learning," *Biomed. Eng. Online*, vol. 22, no. 1, pp. 1–16, 2023, doi: 10.1186/s12938-023-01113-y.

[22] S. Lee *et al.*, "A Survey on Evaluation Metrics for Machine Translation," *Mathematics*, vol. 11, no. 4, pp. 1–22, 2023, doi: 10.3390/math11041006.

[23] S. Saxena, A. Gupta, S. Chauhan, and P. Daniel, "Fuzzy-based Text Augmentation to boost the Statistical Machine Translation for Indic Languages," *Eng. Appl. Artif. Intell.*, vol. 157, no. October 2024, p. 111221, 2025, doi: 10.1016/j.engappai.2025.111221.

[24] E. Novak, L. Bizjak, D. Mladenic, and M. Grobelnik, "Evaluating Text Generation Model Performance by Combining Semantic Meaning and Word Order," *IEEE Access*, vol. 12, no. July, pp. 95265–95277, 2024, doi: 10.1109/ACCESS.2024.3426082.

[25] H. D. Abdulgalil and O. A. Basir, "Next-generation image captioning: A survey of methodologies and emerging challenges from transformers to Multimodal Large Language Models," *Nat. Lang. Process. J.*, vol. 12, no. May, p. 100159, 2025, doi: 10.1016/j.nlp.2025.100159.