Image Captioning Berbasis *Deep Learning* Pada Citra Jantung Anak

Wanda Hamidah
Jurusan Sistem Komputer
Fakultas Ilmu Komputer
Universitas Sriwijaya
Palembang, Indonesia

Firdaus Firdaus
Intelligent Systems Research Group
Fakultas Ilmu Komputer
Universitas Sriwijaya
Palembang, Indonesia
firdaus@unsri.ac.id

Annisa Darmawahyuni
Intelligent Systems Research Group
Fakultas Ilmu Komputer
Universitas Sriwijaya
Palembang, Indonesia
riset.annisadarmawahyuni@gmail.com

Intelligent Systems Research Group
Fakultas Ilmu Komputer
Universitas Sriwijaya
Palembang, Indonesia
wandahamidah90044@gmail.com

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

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

Abstrak—Analisis citra merupakan bidang penting dalam ilmu komputer karena berperan besar dalam pengolahan dan interpretasi data visual secara otomatis. Salah satu area yang banyak dikaji adalah penerapan teknologi *deep learning* untuk mendeteksi pola dan menghasilkan deskripsi dari citra kompleks, termasuk citra jantung anak. Dalam penelitian ini dikembangkan model *image captioning* berbasis *deep learning* untuk menghasilkan deskripsi otomatis dari citra jantung anak. Pendekatan ini menggabungkan arsitektur DenseNet201 sebagai *encoder* untuk mengekstraksi fitur visual dan *Long Short-Term Memory* (LSTM) sebagai *decoder* untuk menghasilkan deskripsi tekstual yang sesuai dengan konteks visual. Proses penelitian meliputi tahap pra-pemrosesan data, ekstraksi fitur, pelatihan model, serta evaluasi menggunakan metrik BLEU, METEOR, dan ROUGE-L. Hasil evaluasi menunjukkan bahwa model yang diusulkan mampu menghasilkan deskripsi yang akurat dan konsisten, dengan skor BLEU sebesar 0,3980, METEOR sebesar 0,6046, dan ROUGE-L sebesar 0,6390 pada data validasi, serta BLEU sebesar 0,4558, METEOR sebesar 0,6816, dan ROUGE-L sebesar 0,6894 pada data *unseen*. Temuan ini menunjukkan bahwa penerapan arsitektur DenseNet201–*LSTM* efektif dalam memahami hubungan antara fitur visual dan representasi tekstual, serta berpotensi meningkatkan efisiensi sistem berbasis visi komputer untuk analisis citra kompleks.

Kata Kunci—Citra Jantung Anak, *Deep Learning, DenseNet201, Image Captioning*, LSTM.

I. PENDAHULUAN

Machine learning dan *deep learning* telah banyak digunakan dalam analisis citra medis karena kemampuannya dalam mengekstraksi pola visual yang kompleks dan mendukung proses diagnosis [1]. Melalui arsitektur seperti *Convolutional Neural Networks* (CNN) dan *Long Short-Term Memory* (LSTM) juga telah diaplikasikan dalam tugas *image captioning* medis, memberikan performa yang impresif dalam menghasilkan deskripsi tekstual dari fitur visual [2]. Penerapan kombinasi kedua arsitektur ini membuka peluang bagi analisis citra medis yang lebih cerdas dan adaptif, khususnya pada domain kardiologi pediatrik.

Meskipun demikian, penerapan *image captioning* pada data medis masih menghadapi sejumlah tantangan, seperti keterbatasan dataset berlabel, kualitas citra yang beragam, serta inkonsistensi anotasi akibat perbedaan interpretasi ahli. Tantangan ini menuntut pengembangan model yang mampu menghasilkan deskripsi akurat secara visual sekaligus relevan secara klinis [3]. *Image captioning* sendiri merupakan teknologi yang menggabungkan *computer vision* dan *natural language processing* untuk menghasilkan deskripsi tekstual dari citra [4]. Dalam konteks pencitraan medis, khususnya citra jantung anak, teknologi ini berpotensi mengotomatisasi interpretasi, meningkatkan konsistensi pelaporan, serta mengurangi beban kerja klinisi[5].

Penyakit jantung bawaan (*congenital heart disease*/CHD) merupakan salah satu anomali janin dan cacat lahir yang paling umum, mempengaruhi sekitar 1% anak di seluruh dunia [6]. Di Indonesia sendiri, angka kejadian penyakit jantung bawaan diperkirakan mencapai sekitar 1% dari total kelahiran bayi, atau sekitar 45.000 kasus setiap tahunnya dengan tingkat keparahan yang bervariasi dari ringan hingga kompleks [7]. Keterbatasan tenaga ahli sonografer dan minimnya ketersediaan data citra jantung anak menjadi tantangan utama dalam diagnosis dini CHD [8]. Deteksi dini kelainan ini sangat penting, namun masih menjadi tantangan karena interpretasi citra jantung membutuhkan keahlian diagnostik khusus [9]. Kompleksitas pola citra medis tersebut menjadikan penerapan *computer vision* dan *deep learning* sebagai pendekatan yang menjanjikan untuk mendukung proses diagnosis yang lebih efisien dan terstandarisasi.

Berbeda dengan penelitian sebelumnya yang hanya berfokus pada klasifikasi citra medis, penelitian ini memperluas pendekatan menuju generasi deskripsi otomatis (*image captioning*) menggunakan arsitektur DenseNet201–LSTM yang diadaptasi khusus untuk citra jantung anak. Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk mengembangkan model *image captioning* pada citra jantung anak dengan memanfaatkan arsitektur DenseNet sebagai *encoder* dan LSTM sebagai *decoder*. *Encoder*

DenseNet digunakan untuk mengekstraksi fitur visual dari citra medis secara efisien, sedangkan *decoder* LSTM bertugas menghasilkan deskripsi tekstual yang sesuai dengan temuan klinis. Pendekatan ini diharapkan dapat meningkatkan efisiensi diagnosis, menstandarkan pelaporan, serta memperluas akses layanan kardiologi pediatrik, khususnya di wilayah dengan keterbatasan tenaga ahli.

II. METODOLOGI

Metodologi penelitian ini menjelaskan tahapan pelaksanaan penelitian secara menyeluruh. Proses dimulai dari pengumpulan data citra jantung anak yang mencakup beberapa kondisi seperti Atrial Septal Defect (ASD), Atrioventricular Septal Defect (AVSD), Ventricular Septal Defect (VSD), dan kasus normal. Selanjutnya dilakukan tahap pra-pemrosesan data, ekstraksi fitur menggunakan DenseNet, serta pelatihan model *image captioning* berbasis DenseNet–LSTM untuk menghasilkan deskripsi otomatis citra medis sesuai konteks klinis. Alur lengkap proses penelitian ditunjukkan pada Gambar 1, yang menggambarkan tahapan pelaksanaan mulai dari akuisisi citra ultrasonografi jantung anak, pra-pemrosesan data, ekstraksi fitur, hingga evaluasi model secara *end-to-end.*

Gambar 1. Diagram Alur Proses *Image Captioning* berbasis DenseNet–LSTM

Data melalui tahap pra-pemrosesan meliputi normalisasi ukuran dan tokenisasi teks, sebelum dilakukan ekstraksi fitur visual menggunakan DenseNet201. Representasi visual yang dihasilkan digabungkan dengan representasi teks melalui *embedding layer* dan diproses oleh LSTM untuk menghasilkan deskripsi otomatis. Evaluasi dilakukan menggunakan metrik BLEU, METEOR, dan ROUGE-L untuk menilai kesesuaian semantik antara caption prediksi dan *ground truth.*

A. *Encoder DenseNet*

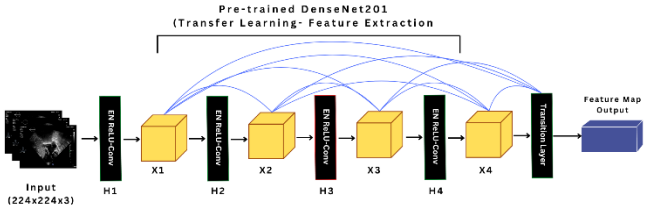
Encoder DenseNet dipilih sebagai ekstraktor fitur visual utama untuk menangkap representasi spasial dan tekstural pada citra jantung anak yang akan digunakan oleh modul *decoder* untuk menghasilkan deskripsi klinis. Arsitektur *DenseNet* menerapkan konektivitas padat antara lapisan sehingga setiap lapisan menerima fitur; hal ini memperkuat *feature reuse*, memperlancar aliran gradien selama pelatihan, dan membantu model mempelajari fitur berlapis dari detail lokal [10]. Dalam penerapan sebagai *encoder* pada tugas captioning medis, beberapa *dense block* disusun berurutan dan dihubungkan oleh *transition layer* yang berfungsi melakukan reduksi dimensi spasial serta kompresi jumlah saluran fitur [11]. Hubungan antarlapisan pada *DenseNet* dijelaskan melalui Persamaan (1), yang menggambarkan aliran informasi dari seluruh lapisan sebelumnya menuju lapisan saat ini.

$$x_\ell = H_\ell([x_0, x_1, \dots, x_{\ell-1}]) \quad (1)$$

Setiap lapisan dalam *dense block* menerima masukan dari semua lapisan sebelumnya melalui operasi *concatenation*, yang memperkuat pemanfaatan ulang fitur dan efisiensi aliran informasi [12]. Lapisan akhir menggunakan *global average pooling* untuk menghasilkan representasi fitur berukuran tetap. Proses ekstraksi fitur tersebut dinyatakan pada Persamaan (2) yang merepresentasikan fungsi pemetaan.

$$F = E_{DenseNet}(I; \Theta) \quad (2)$$

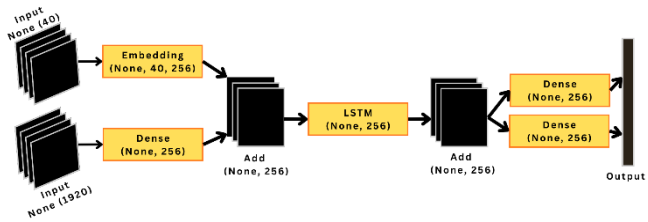
Untuk adaptasi pada citra medis ultrasound/jantung anak, *encoder* di-inisialisasi dengan bobot pra-latih (*transfer learning*) dan di *fine-tune* pada dataset domain spesifik untuk mempercepat konvergensi serta meningkatkan *robustnes* terhadap *noise* dan variasi kontras [13]. Parameter penting yang diatur mencakup *growth rate*, jumlah lapisan per *dense block*, penggunaan *bottleneck* (1×1) untuk efisiensi komputasi, serta regularisasi (*dropout/spatial dropout*). Desain *encoder* ini difokuskan untuk menghasilkan fitur akhir yang informatif dan padat sehingga *decoder* sekuensial dapat menghasilkan caption klinis yang relevan dan konsisten. Ilustrasi arsitektur secara umum ditunjukkan pada Gambar 2.



Gambar 2. Ilustrasi Arsitektur DenseNet201

B. *Decoder LSTM*

Decoder berfungsi untuk menerjemahkan fitur visual hasil ekstraksi dari *encoder* DenseNet menjadi urutan kata yang bermakna dan kontekstual [14]. Pada arsitektur yang digunakan memiliki dua jalur input utama. Jalur pertama berasal dari keluaran *encoder* DenseNet yang kemudian diproyeksikan melalui lapisan Dense dengan 256 neuron untuk menghasilkan representasi visual terkompresi. Jalur kedua berasal dari data teks masukan yang diubah menjadi representasi numerik melalui lapisan embedding berukuran 256 unit. Kedua jalur tersebut kemudian digabungkan melalui *add layer* dan diteruskan ke unit LSTM dengan 256 neuron [15]. Proses ini memungkinkan model menyelaraskan konteks visual dari citra dengan konteks linguistik dari urutan kata dalam satu arsitektur multimodal yang terintegrasi, seperti yang ditampilkan pada Gambar 3.



Gambar 3. Ilustrasi Arsitektur *Embedding* dan LSTM

Fitur gabungan yang dihasilkan dari proses tersebut diproses oleh unit LSTM untuk mempelajari hubungan

temporal antar kata dan mempertahankan konteks visual yang telah diekstraksi sebelumnya. Hasil keluaran dari LSTM kemudian diteruskan ke beberapa lapisan Dense secara berurutan untuk menghasilkan distribusi probabilitas terhadap setiap kata dalam kosakata keluaran [16]. Fungsi kerugian yang digunakan dalam proses pelatihan adalah *categorical cross-entropy loss***,** yang menghitung rata-rata negatif log-probabilitas kata target pada setiap langkah waktu. Optimisasi dilakukan dengan algoritma *Adam* menggunakan *learning rate* adaptif untuk mempercepat konvergensi dan menjaga kestabilan pembelajaran [17]. Untuk mengurangi risiko *overfitting*, diterapkan teknik *dropout* pada koneksi LSTM serta *teacher forcing* sebagian selama proses pelatihan agar model tidak sepenuhnya bergantung pada hasil prediksi sebelumnya [18]. Persamaan (3) menunjukkan formulasi fungsi kerugian yang digunakan dalam proses pelatihan model ini.

$$h_t = LSTM(x_t, h_{t-1}; \Theta) \quad (3)$$

Pendekatan ini memungkinkan *decoder* menghasilkan kalimat deskriptif yang merepresentasikan citra medis secara visual sekaligus bermakna secara klinis. Integrasi antara DenseNet sebagai *encoder* dan LSTM sebagai *decoder* menciptakan *end-to-end* yang mampu memahami keterkaitan antara fitur visual dan bahasa alami, sehingga hasil caption yang dihasilkan lebih akurat dan informatif untuk mendukung interpretasi citra jantung anak.

III. EKSPERIMEN DAN HASIL

Proses eksperimen mencakup persiapan dataset, pelatihan model, serta pengujian menggunakan metrik evaluasi. Selain itu, dilakukan analisis terhadap hasil keluaran model untuk menilai kemampuan dalam menghasilkan deskripsi otomatis yang akurat, konsisten, dan relevan dengan konteks visual citra. Hasil pengujian ini menjadi dasar dalam menilai efektivitas pendekatan yang diusulkan serta potensi penerapannya dalam analisis citra berbasis *deep learning.*

A. Dataset dan *Pre-Processing*

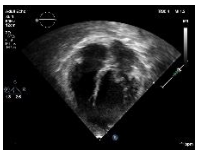
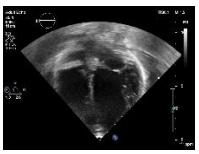
Dataset yang digunakan dalam penelitian ini terdiri atas dua komponen utama, yaitu data citra (*image*) dan data teks (caption)**.** Data citra diperoleh dari hasil pemeriksaan ultrasonografi (USG) jantung anak di Rumah Sakit Moh. Hoesin Palembang selama periode 2020 hingga 2025. Dari setiap pasien diambil beberapa *view* utama seperti *four-chamber* (4CH), *five-chamber* (5CH), *subcostal* (SUB), dan normal masing-masing dipilih tiga frame representatif. Hasilnya diperoleh total 1932 frame, yang terdiri dari 318 frame ASD, 9 frame AVSD, 645 frame VSD, dan 960 frame normal sebagai basis data pelatihan model.

Data caption disusun secara sistematis untuk setiap kategori citra berdasarkan hasil interpretasi visual terhadap citra ultrasonografi jantung anak, mencakup kondisi ASD, VSD, AVSD, dan normal. Setiap citra memiliki satu hingga tiga variasi deskripsi dengan gaya parafrasa untuk memperluas keragaman linguistik serta memastikan konsistensi terminologi anatomi.

Sebaran caption pada data pelatihan terdiri atas 318 caption ASD, 9 AVSD, 645 VSD, dan 960 normal, sedangkan data unseen mencakup 270 pasangan citra dan caption dengan distribusi 64 ASD, 8 AVSD, 198 VSD, dan 100 Normal. Caption pada data unseen mempertahankan pola linguistik yang serupa dengan data pelatihan, namun menggunakan kalimat berbeda untuk menguji kemampuan model dalam menghasilkan deskripsi baru yang tetap akurat dan relevan secara semantik.

Tabel I berikut adalah sampel yang menunjukkan kondisi dan caption:

Tabel I. Sample Data *Image* dan Caption

Image	Caption
	Tidak terlihat lubang pada view SUBCOSTAL. Ruang jantung pada view SUBCOSTAL terdiri dari LA, RA, LV dan RV. kemudian septum terdiri dari IAS (interatrial septum) dan IVS (interventricular septum).
	Lubang terlihat di atrial dekat dengan interatrial septum (IAS) dan posisi lubang juga berada di ruang Left Atrial dan Right Atrial. Gambar usg ini melihat ruang kondisi jantung dari view 4CH. Kondisi ini pasien terkena ASD dilihat dari view 4CH.
	Lubang terlihat di ventrikel dekat dengan interventrikular (IVS) dan posisi lubang juga berada di ruang Left Ventrikel dan Right Ventrikel. Gambar USG ini melihat ruang kondisi jantung dari view 5CH. Pada view 5CH ada bagian Aortic Valve yang dekat dengan hole dan septum IVS. kondisi ini subjek terkena VSD.

Tahap pra-pemrosesan dilakukan untuk menyiapkan data citra dan teks sebelum digunakan dalam pelatihan model *image captioning*. Pada tahap ini, setiap citra jantung yang telah dikumpulkan diolah dan disimpan bersama caption dalam format Excel agar mudah diakses selama proses pelatihan. Proses dimulai dengan menambahkan token khusus "*startseq*" di awal dan "*endseq*" di akhir setiap caption guna menandai batas awal dan akhir kalimat [19]. Selanjutnya dilakukan tokenisasi, yaitu proses mengubah setiap kata menjadi indeks numerik menggunakan *Tokenizer* agar dapat diproses oleh jaringan LSTM [20].

Tahap berikutnya adalah *padding*, yang digunakan untuk menyamakan panjang urutan kata pada seluruh caption sehingga model dapat membaca setiap urutan teks dengan dimensi yang seragam [21]. Selain itu, citra diubah ukurannya menjadi resolusi tetap dan dilakukan normalisasi nilai piksel untuk menstabilkan pelatihan. Seluruh hasil pra-pemrosesan kemudian disimpan dalam folder dataset terstruktur untuk memastikan sinkronisasi antara citra dan teks tetap terjaga.

B. Ekstraksi Fitur

Tahap ekstraksi fitur merupakan lanjutan dari proses pra-pemrosesan, di mana citra yang telah dinormalisasi dan disiapkan digunakan sebagai masukan bagi model *encoder*. Pada penelitian ini, digunakan arsitektur DenseNet201 sebagai model *Convolutional Neural Network* (CNN) untuk mengekstraksi representasi visual dari setiap gambar. Model ini bersifat *pre-trained* pada dataset ImageNet, sehingga mampu mengenali pola visual kompleks yang relevan dengan struktur anatomi jantung [22]. Lapisan akhir DenseNet201 dimodifikasi dengan pengaturan *pooling='avg'* untuk menghasilkan vektor fitur berdimensi 1920 bagi setiap citra.

Seluruh citra berukuran 224×224 piksel kemudian diproses melalui jaringan CNN tersebut, menghasilkan keluaran berupa vektor numerik yang merepresentasikan karakteristik penting dari citra medis. Hasil ekstraksi fitur ini disimpan dalam bentuk *dictionary* dengan pasangan nama file dan vektor fitur, yang selanjutnya digunakan sebagai masukan bagi *decoder* LSTM pada tahap pelatihan. Pendekatan ini memungkinkan model mengenali hubungan antara pola visual jantung dengan deskripsi tekstual secara lebih efisien dan terstruktur.

C. Training dan Testing

Tahap ini dilakukan untuk membagi dataset menjadi beberapa bagian agar model dapat dilatih dan diuji secara objektif. Dari total 1932 gambar beserta caption**,** data dibagi dengan proporsi 80% untuk pelatihan (*training*) dan 20% untuk pengujian (*testing unseen*)**.** Rinciannya meliputi 1572 gambar-caption untuk training, 360 gambar-caption untuk validasi, dan 270 gambar-caption untuk *unseen test*, guna mengevaluasi kemampuan generalisasi model terhadap data baru. Pembagian data tersebut ditunjukkan pada Tabel II, yang menjelaskan proporsi dan jumlah data pada setiap tahap pemrosesan.

Tabel II. Split Data

Training	Test Validation	Test Unseen
1572 gambar caption	360 gambar caption	270 gambar caption

D. Pemodelan LSTM

Pemodelan dilakukan menggunakan arsitektur *Long Short-Term Memory* (LSTM) sebagai *decoder* untuk menghasilkan deskripsi teks berdasarkan fitur visual yang diekstraksi oleh DenseNet201. Model ini menerima masukan berupa vektor fitur citra dan urutan kata dari caption. Lapisan *embedding* digunakan untuk merepresentasikan kata dalam bentuk vektor numerik berdimensi 256, yang kemudian digabungkan dengan fitur visual melalui *add layer* sebelum masuk ke unit LSTM berukuran 256 neuron [23].

Selanjutnya, keluaran dari LSTM diteruskan ke lapisan *Dense* untuk memprediksi kata berikutnya pada setiap langkah waktu hingga membentuk kalimat lengkap. Proses pelatihan dilakukan menggunakan *categorical cross-entropy loss* dengan optimisasi *Adam* dan *learning rate* 0.001. *Dropout* diterapkan untuk mengurangi overfitting, sementara *teacher forcing* digunakan untuk mempercepat konvergensi model [24]. Hasil keluaran dibandingkan dengan caption referensi untuk menghitung akurasi deskripsi dan metrik evaluasi lainnya.

E. Evaluasi Hasil

Evaluasi model dilakukan untuk mengukur sejauh mana *image captioning* mampu menghasilkan deskripsi yang akurat dan sesuai dengan makna klinis pada citra jantung anak. Proses evaluasi dilakukan dengan membandingkan hasil *predicted caption* dengan *ground truth* caption menggunakan tiga metrik utama, yaitu BLEU**,** METEOR, dan ROUGE-L**.** Ketiga metrik ini secara umum menilai kesamaan kata, kesesuaian urutan, serta keterpaduan semantik antara hasil keluaran model dan teks referensi [25].

Validasi ini dilakukan dengan menggunakan Persamaan (4), (5), dan (6), yang merupakan ilustrasi perhitungan tingkat kesesuaian antara hasil prediksi model dengan data *ground truth* yang dijadikan acuan.

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (4)$$

$$METEOR = F_{mean} \times (1 - Penalty) \quad (5)$$

$$ROUGE - L = \frac{(1+\beta^2) \times R_{LCS} \times P_{LCS}}{R_{LCS} + \beta^2 \times P_{LCS}} \quad (6)$$

Berdasarkan hasil pada Tabel III, nilai rata-rata skor menunjukkan performa yang cukup baik, dengan skor BLEU sebesar 0.3980 pada data validasi dan 0.4558 pada data *unseen*. Nilai METEOR dan ROUGE-L yang mencapai lebih dari 0.60 menandakan bahwa model mampu menghasilkan deskripsi yang koheren, relevan secara klinis, dan memiliki kesamaan semantik tinggi terhadap caption referensi.

Tabel III. Hasil Evaluasi Model *Image Captioning*

Metrics	BLEU	METEOR	ROUGE-L
Test Validation	0.3980	0.6046	0.6390
Test Unseen	0.4558	0.6816	0.6894

Hasil keluaran model image captioning menunjukkan kemampuan sistem dalam menghasilkan deskripsi otomatis yang sangat mendekati *ground truth* klinis. Model yang dikembangkan mampu mengenali struktur anatomi dan mendeskripsikan kondisi patologis seperti keberadaan lubang pada septum atrium maupun ventrikel dengan terminologi medis yang tepat. Konsistensi ini mengindikasikan bahwa kombinasi arsitektur DenseNet201 sebagai *feature extractor* dan LSTM sebagai *sequence generator* efektif dalam menangkap hubungan multimodal antara representasi visual dan linguistik pada citra medis.

Selain itu, performa model juga menunjukkan ketahanan terhadap variasi *view* citra seperti 4CH, 5CH, SUB, PSAX, dan Normal. Hasil deskripsi tetap konsisten secara semantik meskipun terdapat perbedaan orientasi dan kompleksitas struktur antar *view*. Dengan demikian, pendekatan yang diusulkan memiliki potensi besar untuk diterapkan dalam sistem bantu diagnosis berbasis kecerdasan buatan yang

mampu meningkatkan efisiensi dan konsistensi interpretasi citra jantung anak. Model juga menunjukkan kemampuan generalisasi yang baik terhadap data baru dengan mempertahankan kesesuaian semantik tinggi antara hasil prediksi dan deskripsi referensi.

Gambar 4 hasil prediksi model pada beberapa citra jantung anak dengan kondisi berbeda menunjukkan bahwa teks berwarna hijau menandakan kata yang sama dengan *ground truth*, mencerminkan tingkat kesesuaian semantik yang tinggi antara hasil prediksi dan deskripsi referensi. Mengindikasikan bahwa model mampu memahami konteks klinis secara mendalam dan menghasilkan deskripsi yang relevan dengan kondisi anatomis yang ditampilkan.

Input	***Ground Truths***	***Predicted Caption***
 30551720250417_AFKAR BOY NURDIANSYAH_20250417180439086_frame_01	lubang terlihat di atrial dekat dengan interatrial septum (IAS) dan posisi lubang juga berada di ruang left atrial dan right atrial. gambar USG ini melihat ruang kondisi jantung dari view subcostal. kondisi ini pasien terkena ASD dilihat dari view SUBCOSTAL.	lubang terlihat di atrial dekat dengan interatrial septum (IAS) dan posisi lubang juga berada di ruang left atrial dan right atrial. gambar usg ini melihat ruang kondisi jantung dari view SUBCOSTAL. kondisi ini pasien terkena ASD dilihat dari view SUBCOSTAL.
 48030620211115_M.OMAR_20211115061203093 _frame_02	view PSAX sedikit agak berbeda dengan view lain adapun ruang jantung pada view PSAX adalah RVOT, left atrial, right atrial, pulmonary artery dan aortic valve. untuk septum juga penamaan sedikit berbeda posisi lubang didekat sepltu triscupd valve dan pulmonic valve. gambar USG ini melihat ruang kondisi jantung dari view PSAX.	view PSAX memiliki sedikit perbedaan dibandingkan view lainnya. ruang-ruang jantung yang terlihat pada view PSAX meliputi RVOT, left atrial, right atrial, pulmonary artery, dan aortic valve. penamaan pada septum juga sedikit berbeda, dengan lubang terletak di dekat tricuspid valve dan pulmonic valve. gambar USG ini menunjukkan kondisi ruang jantung dari view PSAX

Gambar 4. Contoh hasil evaluasi model; teks berwarna hijau menunjukkan kata yang identik dengan *ground truth.*

IV. DISKUSI

Dalam penelitian ini telah mengeksplorasi penerapan *deep learning* untuk menghasilkan deskripsi otomatis citra jantung anak melalui metode *image captioning* berbasis DenseNet201–LSTM. Pendekatan ini memanfaatkan kekuatan *encoder–decoder* dalam memahami hubungan antara pola visual dan konteks klinis. Pada penelitian sebelumnya, metode berbasis CNN umumnya digunakan untuk klasifikasi atau segmentasi citra medis, namun penelitian ini memperluas penerapannya hingga ke level deskriptif melalui generasi teks otomatis. Model yang diusulkan mampu mengenali struktur anatomi jantung dan menghasilkan caption yang relevan dengan temuan klinis, baik pada data *validation* maupun *unseen*.

Kinerja model menunjukkan peningkatan yang signifikan dibanding metode deskriptif konvensional, ditunjukkan oleh nilai metrik BLEU, METEOR, dan ROUGE-L yang cukup tinggi. Hasil ini membuktikan bahwa kombinasi DenseNet sebagai *feature extractor* dan LSTM sebagai *sequence generator* efektif dalam menangkap korelasi multimodal antara gambar dan teks. Pendekatan ini tidak hanya berpotensi mendukung efisiensi interpretasi citra medis, tetapi juga membuka peluang integrasi otomatis dalam pelaporan hasil USG jantung anak secara konsisten dan cepat.

V. KESIMPULAN

Penelitian ini berhasil mengembangkan model *image captioning* berbasis DenseNet201–LSTM untuk menghasilkan deskripsi otomatis pada citra jantung anak. Hasil evaluasi menunjukkan bahwa model mampu memahami hubungan antara fitur visual dan konteks klinis dengan baik, yang ditunjukkan oleh skor BLEU, METEOR, dan ROUGE-L yang cukup tinggi pada data *validation* maupun *unseen*. Pendekatan ini berpotensi membantu proses interpretasi citra medis secara otomatis, meningkatkan konsistensi pelaporan, serta mendukung efisiensi diagnosis.

Untuk penelitian selanjutnya, mekanisme *attention* dapat diintegrasikan ke dalam arsitektur model guna meningkatkan kesesuaian semantik antara fitur visual dan deskripsi klinis, sehingga hasil *caption* yang dihasilkan menjadi lebih akurat dan kontekstual.

REFERENSI

[1] J. Zhou, M. Du, S. Chang, and Z. Chen, “Artificial intelligence in echocardiography: detection, functional evaluation, and disease diagnosis,” Dec. 01, 2021, *BioMed Central Ltd.* doi: 10.1186/s12947-021-00261-2.

[2] P. Balasundaram, K. Swaminathan, O. Sampath, and P. Km, “Concept Detection and Caption Prediction of Radiology Images Using Convolutional Neural Networks Keywords,” 2024. [Online]. Available: https://github.com/pradeepkmaran/imageclef2024

[3] A. Selivanov, O. Y. Rogov, D. Chesakov, A. Shelmanov, I. Fedulova, and D. V. Dylov, “Medical image captioning via generative pretrained

transformers," *Sci Rep*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-31223-5.

[4] S. Park and J. Paik, "RefCap: image captioning with referent objects attributes," *Sci Rep*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-48916-6.

[5] A. Wang, T. T. Doan, C. Reddy, and P. N. Jone, "Artificial Intelligence in Fetal and Pediatric Echocardiography," Jan. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/children12010014.

[6] M. Zubrzycki, R. Schramm, A. Costard-Jäckle, J. Grohmann, J. F. Gummert, and M. Zubrzycka, "Cardiac Development and Factors Influencing the Development of Congenital Heart Defects (CHDs): Part I," Jul. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/ijms25137117.

[7] I. K. Murni *et al.*, "Delayed diagnosis in children with congenital heart disease: a mixed-method study," *BMC Pediatr*, vol. 21, no. 1, Dec. 2021, doi: 10.1186/s12887-021-02667-3.

[8] M. N. Rachmatullah, S. Nurmaini, A. I. Sapitri, A. Darmawahyuni, B. Tutuko, and Firdaus, "Convolutional neural network for semantic segmentation of fetal echocardiography based on four-chamber view," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 4, pp. 1987–1996, Aug. 2021, doi: 10.11591/EEI.V10I4.3060.

[9] X. Jiang, J. Yu, J. Ye, W. Jia, W. Xu, and Q. Shu, "A deep learning-based method for pediatric congenital heart disease detection with seven standard views in echocardiography," *World Journal of Pediatric Surgery*, vol. 6, no. 3, Jun. 2023, doi: 10.1136/wjps-2023-000580.

[10] T. Zhou, X. Ye, H. Lu, X. Zheng, S. Qiu, and Y. Liu, "Dense Convolutional Network and Its Application in Medical Image Analysis," 2022, *Hindawi Limited*. doi: 10.1155/2022/2384830.

[11] P. Podder, F. B. Alam, M. R. H. Mondal, M. J. Hasan, A. Rohan, and S. Bharati, "Rethinking Densely Connected Convolutional Networks for Diagnosing Infectious Diseases," *Computers*, vol. 12, no. 5, May 2023, doi: 10.3390/computers12050095.

[12] L. Tan, Q. Fu, and J. Li, "An Improved Neural Network Model Based on DenseNet for Fabric Texture Recognition," *Sensors*, vol. 24, no. 23, Dec. 2024, doi: 10.3390/s24237758.

[13] M. R. Hosseinzadeh Taher, F. Haghighi, M. B. Gotway, and J. Liang, "Large-scale benchmarking and boosting transfer learning for medical image analysis," *Med Image Anal*, vol. 102, May 2025, doi: 10.1016/j.media.2025.103487.

[14] A. A. E. Osman, M. A. W. Shalaby, M. M. Soliman, and K. M. Elsayed, "Novel concept-based image captioning models using LSTM and multi-encoder transformer architecture," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-69664-1.

[15] T. Bai, S. Zhou, Y. Pang, J. Luo, H. Wang, and Y. Du, "An image caption model based on attention mechanism and deep reinforcement learning," *Front Neurosci*, vol. 17, 2023, doi: 10.3389/fnins.2023.1270850.

[16] F. Hoseini and A. Y. Notash, "Image captioning using bidirectional LSTM neural network," *Discover Artificial Intelligence*, vol. 5, no. 1, Dec. 2025, doi: 10.1007/s44163-025-00315-8.

[17] Y. L. Peng and W. P. Lee, "Practical guidelines for resolving the loss divergence caused by the root-mean-squared propagation optimizer," *Appl Soft Comput*, vol. 153, Mar. 2024, doi: 10.1016/j.asoc.2024.111335.

[18] P. Ravinder and S. Srinivasan, "Automated Medical Image Captioning with Soft Attention-Based LSTM Model Utilizing YOLOv4 Algorithm," *Journal of Computer Science*, vol. 20, no. 1, pp. 52–68, 2024, doi: 10.3844/jcssp.2024.52.68.

[19] S. Tyagi *et al.*, "Novel Advance Image Caption Generation Utilizing Vision Transformer and Generative Adversarial Networks," *Computers*, vol. 13, no. 12, Dec. 2024, doi: 10.3390/computers13120305.

[20] H. Deng and A. Alkhayyat, "Sentiment analysis using long short term memory and amended dwarf mongoose optimization algorithm," *Sci Rep*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-01834-1.

[21] M. Inayathulla, "Image Caption Generation using Deep Learning For Video Summarization Applications," 2024. [Online]. Available: www.ijacsa.thesai.org

[22] G. Ayana, K. Dese, A. M. Abagaro, K. C. Jeong, S. Do Yoon, and S. W. Choe, "Multistage transfer learning for medical images," *Artif Intell Rev*, vol. 57, no. 9, Sep. 2024, doi: 10.1007/s10462-024-10855-7.

[23] P. V. Kavitha and V. Karpagam, "Image captioning deep learning model using ResNet50 encoder and hybrid LSTM–GRU decoder optimized with beam search," *Automatika*, vol. 66, no. 3, pp. 394–410, 2025, doi: 10.1080/00051144.2025.2485695.

[24] F. Lv, R. Wang, L. Jing, and P. Dai, "HIST: Hierarchical and sequential transformer for image captioning," *IET Computer Vision*, vol. 18, no. 7, pp. 1043–1056, Oct. 2024, doi: 10.1049/cvi2.12305.

[25] A. Thobhani *et al.*, "A novel image captioning model with visual-semantic similarities and visual representations re-weighting," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 7, Sep. 2024, doi: 10.1016/j.jksuci.2024.102127.