Syntax:
Journal of Software Engineering, Computer Science and Information Technology
p-ISSN: 2776-7027, e-ISSN: 2723-0538
Volume: 6, Nomor: 1, Juni 2025

IMPLEMENTASI CONVOLUTIONAL NEURAL NETWORK (CNN) DAN CONTRASTIVE LANGUAGE-IMAGE PRETRAINING (CLIP) UNTUK PREDIKSI GENRE FILM BERBASIS ANALISIS POSTER

Sebastian Kurniawan Windu Wiwaha[1], Kartono Pinaryanto[2]

[1,2] Unversitas Sanata Dharma
Paingan, Maguwoharjo, Depok, Sleman, Yogyakarta 55281

[1]sebastiank27sept@gmail.com, [2]kartono@usd.ac.id

ABSTRAK

***Abstrak*— Industri perfilman terus berkembang pesat, menghasilkan ribuan film setiap tahun. Klasifikasi genre film menjadi krusial untuk pengelompokan dan sistem rekomendasi. Poster film, sebagai elemen visual utama, seringkali merepresentasikan genre melalui objek, warna, dan desain, namun informasi tekstual seperti plot juga signifikan. Penelitian ini bertujuan membandingkan performa *Convolutional Neural Network* (CNN) dan *Contrastive Language-Image Pretraining* (CLIP) dalam klasifikasi genre film multi-label menggunakan analisis poster dan plot. Dataset dari IMDb dan OMDb diproses melalui tahap *preprocessing*. Model CNN menggunakan arsitektur BiT-ResNet50, sementara CLIP menggunakan ViT-B/16, ViT-L/14, dan RN50x16 untuk poster, serta BERT untuk analisis plot. Eksperimen melibatkan variasi *batch size*, *learning rate*, dan *optimizer*. Hasil menunjukkan CLIP (ViT-L/14) lebih unggul dengan akurasi 83,2% dan *Hamming Loss* 0,1678, dibandingkan CNN dengan akurasi 77,9%. Integrasi analisis plot menggunakan BERT meningkatkan akurasi sekitar 5% dibandingkan metode berbasis poster saja. Studi ini membuktikan bahwa kombinasi model *vision-language* (CLIP) dan analisis teks (BERT) lebih efektif daripada CNN konvensional untuk klasifikasi genre film.**

***Kata Kunci*—klasifikasi genre film, CNN, CLIP, deep learning, poster film, multi label classification.**

ABSTRACT

Abstract— **The film industry continues to develop rapidly, producing thousands of films annually. Film genre classification has become crucial for categorization and recommendation systems. Film posters, as primary visual elements, often represent genres through objects, colors, and design, while textual information such as plot is equally significant. This research aims to compare the performance of Convolutional Neural Network (CNN) and Contrastive Language-Image Pretraining (CLIP) in multi-label film genre classification using poster and plot analysis. The dataset from IMDb and OMDb was processed through preprocessing stages. The CNN model used BiT-ResNet50 architecture, while CLIP used ViT-B/16, ViT-L/14, and RN50x16 for posters, along with BERT for plot analysis. Experiments involved variations in batch size, learning rate, and optimizer. Results show CLIP (ViT-L/14) outperformed with 83.2% accuracy and Hamming Loss of 0.1678, compared to CNN with 77.9% accuracy. Integrating plot analysis using BERT improved accuracy by approximately 5% compared to poster-only methods. This study demonstrates that the combination of vision-language models (CLIP) and text analysis (BERT) is more effective than conventional CNN for film genre classification.**

Keywords—film genre classification, CNN, CLIP, deep learning, movie posters, multi-label classification.

Syntax:
Journal of Software Engineering, Computer Science and Information Technology
p-ISSN: 2776-7027, e-ISSN: 2723-0538
Volume: 6, Nomor: 1, Juni 2025

I. PENDAHULUAN

Industri perfilman global mengalami pertumbuhan signifikan, dengan ribuan film baru dirilis setiap tahun. Di tengah volume konten yang masif ini, genre film berfungsi sebagai penanda identitas utama yang tidak hanya membantu dalam kategorisasi tetapi juga krusial dalam memandu pilihan penonton. Akurasi identifikasi genre menjadi sangat penting untuk memenuhi ekspektasi audiens dan mengoptimalkan pengalaman menonton. Poster film, sebagai garda depan pemasaran visual, memegang peranan penting dalam mengkomunikasikan esensi sebuah film. Lebih dari sekadar alat promosi, poster seringkali menjadi representasi visual dari genre melalui penggunaan elemen desain spesifik seperti skema warna, tipografi, komposisi objek, dan gaya visual keseluruhan. Poster yang efektif dapat memberikan petunjuk mengenai alur cerita, suasana, dan tema, sekaligus menarik perhatian calon penonton. Oleh karena itu, analisis poster untuk klasifikasi genre otomatis menjadi area penelitian yang relevan dan menjanjikan[1].

Perkembangan teknologi *deep learning*, khususnya *Convolutional Neural Networks* (CNN), telah mendorong banyak penelitian dalam klasifikasi genre film otomatis berbasis poster. CNN terbukti efektif dalam mengekstraksi fitur visual hierarkis dari gambar, mulai dari fitur dasar seperti tepi dan warna hingga fitur kompleks seperti objek dan tekstur yang dapat berkorelasi dengan genre tertentu. Beberapa penelitian sebelumnya telah menunjukkan potensi CNN dalam tugas ini yaitu:

Penelitian pertama oleh Chu dan Guo[2] menerapkan *deep neural network* untuk klasifikasi genre film berbasis poster menggunakan dataset 8.191 poster film Hollywood (1980-2015) dengan 23 kategori genre. Meskipun penelitian ini mengintegrasikan ekstraksi visual dan deteksi objek dengan thresholding adaptif untuk klasifikasi multi-label, akurasi yang dicapai hanya 18,73%. Hasil ini mengindikasikan tingginya kompleksitas dalam tugas klasifikasi genre film dan menunjukkan perlunya pendekatan yang lebih mendalam.

Penelitian kedua dilakukan oleh Hossain[3] yang mengimplementasikan berbagai arsitektur CNN konvensional seperti VGG16, ResNet50, dan InceptionV3. Penelitian ini memberikan perbandingan sistematis antar arsitektur CNN dan berhasil menunjukkan peningkatan performa dibandingkan penelitian sebelumnya dengan akurasi 91,15%. Namun, pendekatan ini masih menghadapi keterbatasan dalam menangani kompleksitas dan variabilitas visual poster film yang tinggi, terutama dalam skenario *multi-label classification*.

Penelitian yang dilakukan oleh Barney dan Kaya[4] mengimplementasikan model ResNet34 untuk deteksi genre film berdasarkan poster, namun penelitian ini menghadapi tantangan signifikan dalam mencapai performa yang optimal. Hasil yang kurang maksimal terhadap beberapa genre mengindikasikan keterbatasan model dalam memfokuskan pembelajaran pada karakteristik visual spesifik poster film, yang berbeda dengan domain gambar natural pada umumnya. Penelitian ini menyoroti pentingnya spesialisasi model untuk domain poster film yang memiliki karakteristik visual unik.

Penelitian selanjutnya adalah Wi, Jang, dan Kim[5] memperkenalkan pendekatan transfer learning dengan memanfaatkan model pretrained seperti ResNet dan MobileNet, dengan fokus pada optimalisasi *trade-off* antara akurasi dan kecepatan inferensi untuk aplikasi *real-time*. Kontribusi signifikan mereka adalah pengembangan *inter-channel features* yang mampu menangkap korelasi antar *channel* warna dalam poster film, memberikan representasi visual yang lebih kaya untuk proses klasifikasi.

Paradigma baru dalam *computer vision* diperkenalkan oleh Radford[6] melalui pengembangan CLIP (Contrastive Language-Image Pretraining). Model ini dilatih pada dataset masif 400 juta pasangan gambar-teks dan membuktikan kemampuan transfer learning yang superior pada lebih dari 30 dataset visi komputer. Kemampuan fundamental CLIP dalam memahami korelasi semantik antara modalitas visual dan tekstual membuka peluang baru yang menjanjikan untuk aplikasi klasifikasi genre film yang lebih akurat.

Validasi lebih lanjut terhadap potensi CLIP dilakukan oleh Shen[7] yang mengeksplorasi aplikasinya pada berbagai tugas vision-and-language kompleks seperti Visual Question Answering (VQA), Image Captioning, dan Vision-and-Language Navigation (VLN). Penelitian ini memvalidasi kemampuan CLIP sebagai encoder visual yang powerful untuk tugas-tugas yang memerlukan pemahaman lintas modalitas, memberikan dasar yang kuat untuk eksplorasi lebih lanjut dalam domain klasifikasi genre film.

Meskipun berbagai pendekatan telah dikembangkan, analisis literatur mengungkapkan beberapa gap signifikan yang belum terakomodasi sebagai berikut:

Pertama, belum ada penelitian yang secara sistematis membandingkan performa CNN konvensional dengan model *vision-language* modern seperti CLIP dalam konteks spesifik klasifikasi genre film. Penelitian-penelitian terdahulu cenderung fokus pada satu pendekatan tanpa melakukan evaluasi komparatif yang komprehensif, sehingga sulit untuk menentukan pendekatan optimal.

Syntax:
Journal of Software Engineering, Computer Science and Information Technology
p-ISSN: 2776-7027, e-ISSN: 2723-0538
Volume: 6, Nomor: 1, Juni 2025

Kedua, sebagian besar penelitian hanya memanfaatkan informasi visual dari poster tanpa mengintegrasikan informasi tekstual seperti plot atau sinopsis film. Padahal, informasi tekstual dapat memberikan konteks semantik tambahan yang berpotensi meningkatkan akurasi klasifikasi secara signifikan, terutama dalam kasus di mana visual poster tidak cukup representatif terhadap genre sebenarnya.

Ketiga, evaluasi performa dalam penelitian-penelitian terdahulu masih menunjukkan akurasi yang relatif rendah pada penelitian pertama[2], mengindikasikan perlunya eksplorasi pendekatan yang lebih *advanced* dengan pemanfaatan dataset yang lebih representatif dan teknologi yang lebih mutakhir.

Keempat, tidak ada penelitian yang mengeksplorasi potensi integrasi multi-modal antara analisis poster (visual) dan analisis plot (tekstual) menggunakan *framework unified* seperti CLIP yang dapat memproses kedua modalitas secara simultan dan sinergis.

Kelima, kurangnya eksplorasi sistematis terhadap berbagai konfigurasi *hyperparameter* dan arsitektur model dalam konteks spesifik klasifikasi genre film, yang dapat memberikan insights mendalam tentang faktor-faktor yang mempengaruhi performa model.

Keenam, penelitian[4] menunjukkan bahwa penggunaan arsitektur CNN standar seperti ResNet34 tanpa adaptasi khusus untuk domain poster film menghasilkan performa yang suboptimal, terutama untuk genre-genre tertentu. Hal ini mengindikasikan perlunya pendekatan yang lebih spesifik dan terfokus pada karakteristik unik poster film sebagai domain visual yang berbeda dari gambar *natural* pada umumnya.

Identifikasi gap-gap ini menjadi semakin relevan mengingat evaluasi desain poster di industri masih sering dilakukan secara manual oleh rumah produksi atau desainer. Proses manual ini memiliki keterbatasan yang signifikan, termasuk memakan waktu lama, biaya tinggi, dan potensi inkonsistensi akibat subjektivitas penilai. Kebutuhan akan metode otomatis yang cepat, konsisten, dan akurat untuk mengidentifikasi genre dari poster menjadi semakin mendesak, terutama untuk memastikan keselarasan antara representasi visual poster dengan genre film yang sebenarnya. Sistem klasifikasi genre otomatis yang andal dapat membantu profesional industri dalam mengevaluasi efektivitas poster dan menyusun strategi pemasaran yang lebih tepat sasaran.

Kemajuan terbaru dalam *multi-modal learning*, khususnya model seperti *Contrastive Language-Image Pretraining* (CLIP), menawarkan pendekatan baru yang menjanjikan. CLIP dilatih untuk memahami hubungan semantik antara data visual (gambar) dan data tekstual (deskripsi). Kemampuan ini memungkinkan CLIP untuk memetakan representasi poster film dan deskripsi genre ke dalam ruang vektor bersama, sehingga dapat meningkatkan pemahaman kontekstual dan akurasi klasifikasi, terutama untuk tugas *multi-label classification* yang kompleks, [6]. Integrasi antara analisis visual poster dengan informasi tekstual, seperti plot film, menggunakan model seperti BERT (*Bidirectional Encoder Representations from Transformers*), berpotensi lebih lanjut memperkaya representasi fitur dan meningkatkan performa klasifikasi.

Berikut ini adalah tabel 1 hasil rangkuman literature review.

Tabel 1
Rangkuman Literature Review

Ref & Peneliti	Metode, Dataset dan Hasil	Kontribusi Utama dan Keterbatasan
[2] Chu dan Guo	**Metode** Deep Neural Network + Ekstraksi visual + Deteksi objek + Thresholding adaptif **Dataset** 8.191 poster film Hollywood dengan 23 kategori genre. **Hasil** akurasi 18,73%.	**Kontribusi utama:** • Penelitian pionir DNN untuk genre film • Integrasi ekstraksi visual dan deteksi objek • Multi-label classification **Keterbatasn:** • Akurasi sangat rendah • Kompleksitas tinggi tugas klasifikasi genre film
[3] Hossain	**Metode** CNN Konvensional • VGG16 • ResNet50 • InceptionV3 **Dataset** besar. **Hasil** akurasi 91,15%.	**Kontribusi utama:** • Perbandingan sistematis antar arsitektur CNN • Peningkatan performa dibanding penelitian sebelumnya **Keterbatasan:** • Kesulitan menangani kompleksitas visual poster • Keterbatasan dalam multi-label classification
[4] Barney dan Kaya	**Metode** ResNet34 **Dataset** berupa poster film. **Hasil** kurang maksimal untuk beberapa genre	**Kontribusi utama:** • Menyoroti pentingnya spesialisasi model • Identifikasi karakteristik unik poster film **Keterbatasan:** • Model tidak fokus pada karakteristik visual spesifik poster • Perbedaan dengan domain gambar natural
[5] Wi, Jang, dan Kim	**Metode** Transfer Learning • ResNet (pretrained) • MobileNet (pretrained) + Inter-channel features **Dataset** berupa poster film.	**Kontribusi utama:** • Transfer learning untuk real-time application • Inter-channel features untuk korelasi warna • Representasi visual yang lebih kaya **Keterbatasan:** • Fokus pada kecepatan mungkin mengurangi akurasi maksimal

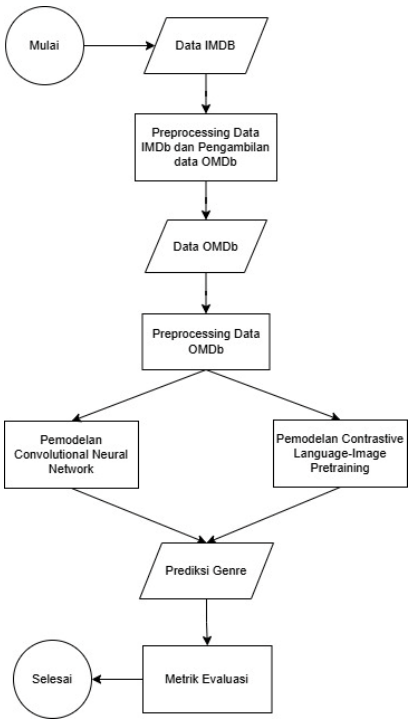
Syntax:
Journal of Software Engineering, Computer Science and Information Technology
p-ISSN: 2776-7027, e-ISSN: 2723-0538
Volume: 6, Nomor: 1, Juni 2025

	Hasil optimalisasi trade-off akurasi vs kecepatan	
[6] Radford	**Metode** CLIP **Dataset** 400 juta pasangan gambar-teks 30+ dataset visi computer **Hasil** superior transfer learning performance	**Kontribusi utama:** • Paradigma baru computer vision • Korelasi semantic visual-tekstual • Transfer learning superior • Membuka peluang baru klasifikasi genre **Keterbatasan:** • Memerlukan dataset besar • Kompleksitas komputasi tinggi
[7] Shen	**Metode** CLIP untuk tugas Vision-Language • VQA • Image Captioning • VLN **Dataset** berbagai dataset untuk tugas vision language **Hasil** validasi kemampuan CLIP sebagai encoder visual yang powerful	**Kontribusi utama:** • Validasi CLIP untuk pemahaman lintas modalitas • Dasar kuat untuk eksplorasi klasifikasi genre film **Keterbatasan:** • Belum spesifik untuk domain poster film • Perlu adaptasi untuk genre classification

Berdasarkan hasil tabel 1 literatur review dan pemaparan gap yang telah diidentifikasi, penelitian ini bertujuan untuk mengimplementasikan dan membandingkan secara sistematis performa arsitektur CNN modern (menggunakan *Transfer Learning* dengan BiT-ResNet50) dan model CLIP (*multi-modal*) dalam tugas klasifikasi genre film *multi-label* berbasis analisis poster dan plot film. Dengan memanfaatkan dataset yang dikumpulkan dari IMDb dan OMDb, penelitian ini menguji berbagai konfigurasi *hyperparameter* untuk mengidentifikasi pendekatan optimal. Hasil penelitian ini diharapkan dapat memberikan kontribusi pada pengembangan sistem klasifikasi genre yang lebih akurat dan efisien, serta memberikan wawasan tentang keunggulan pendekatan *multi-modal* dibandingkan metode berbasis visual murni dalam konteks analisis konten media.

II. METODOLOGI PENELITIAN

Penelitian ini dirancang untuk mengembangkan dan mengevaluasi sistem klasifikasi genre film *multi-label* menggunakan pendekatan *deep learning* berbasis analisis poster dan plot film. Alur kerja penelitian, seperti diilustrasikan pada Gambar 1, mencakup beberapa tahapan utama: akuisisi data, *preprocessing*, ekstraksi fitur dan pemodelan menggunakan CNN dan CLIP, serta evaluasi performa.



Gambar. 1 Alur Proses Penelitian

A. Pengumpulan Data

Tahap pengumpulan data bertujuan mengumpulkan dataset komprehensif yang berisi poster film beserta metadata relevan, termasuk genre dan plot. Dataset penelitian ini dibangun melalui integrasi dua sumber utama: IMDb (*Internet Movie Database*) dan OMDb (*Open Movie Database*). Pemilihan strategi integrasi ini didasarkan pada karakteristik komplementer kedua sumber data, di mana IMDb menyediakan metadata struktural yang komprehensif namun terbatas dalam penyediaan aset visual, sementara OMDb menyediakan akses yang lebih mudah terhadap URL poster film dan informasi plot yang detail.

1. **Pengambilan Data Primer IMDb**: Metadata awal diperoleh dari dataset publik IMDb Developer (https://datasets.imdbws.com/) yang menyediakan akses terhadap informasi jutaan entri film, acara TV, dan konten entertainment lainnya. Dataset IMDb dipilih sebagai sumber primer karena menyediakan sistem identifikasi unik (tconst) yang konsisten dan metadata genre yang telah tervalidasi oleh komunitas global.

   Data mentah IMDb mencakup atribut lengkap berupa tconst (ID unik IMDb), titleType, primaryTitle, originalTitle, startYear, endYear, runtimeMinutes, dan genres. Namun, setelah melalui proses penyaringan ketat (dijelaskan detail di bagian Preprocessing), hanya atribut

esensial yang dipertahankan meliputi tconst (ID unik IMDb), titleType, dan genres. Penyaringan ini dilakukan untuk memfokuskan dataset pada film dengan informasi genre yang valid dan menghilangkan entri yang tidak relevan seperti acara TV, dokumenter, atau film dengan metadata yang tidak lengkap.

Meskipun IMDb menyediakan metadata yang komprehensif, dataset publik yang tersedia tidak menyertakan URL poster film atau akses langsung terhadap aset visual. Keterbatasan ini menjadi tantangan signifikan mengingat penelitian ini memerlukan analisis visual poster sebagai komponen utama klasifikasi genre.

2. **Pengambilan Data Sekunder OMDb**: Untuk mengatasi keterbatasan data IMDb, dilakukan pengayaan data melalui OMDb API (*https://www.omdbapi.com/*) yang menyediakan akses terhadap URL poster film dan informasi plot yang detail. OMDb dipilih sebagai sumber sekunder karena kemampuannya dalam menyediakan metadata visual yang kompatibel dengan ID unik IMDb (tconst), memungkinkan integrasi yang sesuai antara kedua sumber data.

   Proses pengambilan data OMDb dilakukan melalui web crawling sistematis menggunakan tconst dari data IMDb yang telah difilter. Untuk setiap ID film, sistem melakukan query ke OMDb API untuk memperoleh informasi tambahan yang mencakup Title, Year, Rated, Released, Runtime, Genre, Director, Writer, Actors, Plot, Language, Country, Awards, Poster (URL), Ratings, Metascore, imdbRating, imdbVotes, imdbID, Type, DVD, BoxOffice, Production, Website, dan Response.

   Kriteria Seleksi Data: Dari seluruh informasi yang tersedia di OMDb, penelitian ini memfokuskan pada dua atribut kritis: URL poster film dan ringkasan plot. Hanya film yang memenuhi kriteria validasi berikut yang dipertahankan dalam dataset final: (1) memiliki URL poster yang valid dan dapat diakses, (2) menyediakan informasi plot yang memadai, dan (3) konsistensi genre antara data IMDb dan OMDb.

B. Preprocessing Data

Preprocessing merupakan langkah krusial untuk memastikan kualitas dan kesesuaian data sebelum digunakan dalam pelatihan model. Proses ini dilakukan dalam dua fase, sesuai dengan sumber data:

1. **Preprocessing Data IMDb**: Dataset awal IMDb (>11 juta entri) disaring secara ekstensif. Kriteria penyaringan meliputi: mempertahankan hanya titleType='movie'; menghapus entri dengan data kosong/tidak lengkap; mengecualikan film dewasa (isAdult=1); menghapus duplikasi judul; membatasi genre pada 10 kategori teratas (Action, Adventure, Comedy, Crime, Drama, Family, Horror, Mystery, Sci-Fi, Thriller – berdasarkan analisis frekuensi awal, meskipun daftar akhir bisa sedikit berbeda tergantung data OMDb); dan membatasi tahun rilis (startYear >= 2000). Setelah tahap ini, tersisa 18.669 kandidat film.

2. **Preprocessing Data OMDb**: Data hasil *crawling* dari OMDb (berdasarkan 18.669 ID IMDb) dibersihkan lebih lanjut. Film tanpa URL poster yang valid atau tanpa informasi plot dihapus. Film yang genre-nya di OMDb tidak termasuk dalam 10 kategori utama yang telah ditentukan juga dieliminasi. Proses ini menghasilkan dataset final sebanyak 15.150 film yang digunakan untuk pelatihan dan evaluasi.

3. **Transformasi Data**:
   a. *Normalisasi Gambar*: Semua gambar poster diubah ukurannya menjadi 224x224 piksel, sesuai dengan input standar untuk model BiT dan CLIP. Nilai piksel dinormalisasi menggunakan rata-rata dan standar deviasi dari dataset ImageNet untuk model BiT, atau normalisasi spesifik CLIP untuk model CLIP.
   b. *Encoding Genre*: Label genre, yang awalnya berupa string (misal "Action,Adventure,Sci-Fi"), diubah menjadi format *multi-label binary encoding*. Setiap film direpresentasikan oleh vektor biner berukuran N (jumlah genre target, N=10), di mana nilai '1' menunjukkan keberadaan genre tersebut dan '0' sebaliknya. Ini sesuai untuk *multi-label classification*.
   c. *Tokenisasi Teks*: Untuk model CLIP yang melibatkan analisis plot, teks plot film diproses menggunakan *tokenizer* spesifik CLIP untuk mengubah teks menjadi urutan token numerik yang dapat diproses oleh *text encoder*.

C. Pembagian Data

Dataset final (15.150 film) dibagi menjadi tiga subset untuk proses pelatihan, validasi, dan pengujian model berdasarkan praktik pembagian data standar dalam machine learning[8] sebagai berikut:

a. **Training Set**: 70% (10.605 sampel). Digunakan untuk melatih parameter model CNN dan CLIP.

Syntax:
Journal of Software Engineering, Computer Science and Information Technology
p-ISSN: 2776-7027, e-ISSN: 2723-0538
Volume: 6, Nomor: 1, Juni 2025

b. **Validation Set**: 15% (2.272 sampel). Digunakan selama pelatihan untuk *tuning hyperparameter* dan memantau *overfitting* (misalnya, untuk *early stopping* dan penyesuaian *learning rate*).
c. **Testing Set**: 15% (2.273 sampel). Digunakan untuk evaluasi akhir performa model yang telah dilatih pada data yang belum pernah dilihat sebelumnya. Pembagian dilakukan secara acak namun memastikan stratifikasi genre (jika memungkinkan) untuk menjaga distribusi genre yang seimbang di setiap subset.

D. Pemodellan (Modelling)

Penelitian ini mengimplementasikan dan membandingkan dua arsitektur *deep learning* utama:

1. **Model CNN (BiT-ResNet50):**
   a. *Arsitektur*: Menggunakan model *pre-trained* BiT-M (berbasis ResNet50) yang dilatih pada dataset skala besar (misalnya ImageNet-21k). Pendekatan *Transfer Learning* ini memanfaatkan fitur visual kaya yang telah dipelajari oleh BiT.
   b. *Struktur*: *Base model* BiT-ResNet50 berfungsi sebagai *feature extractor*. Lapisan atas (*top layers*) diganti dengan lapisan *Global Average Pooling*, diikuti oleh *Dense layer* (256 neuron, aktivasi ReLU), *Batch Normalization*, *Dropout* (rate 0.5), dan *output layer* berupa *Dense layer* dengan 10 neuron (sesuai jumlah genre target) dan fungsi aktivasi *Sigmoid* untuk menangani *multi-label classification*. Struktur model dapat dilihat pada Tabel 2.

Tabel 2
Struktur Model BIT

Layer Name	Type	Kernel Size / Stride	*Output* Shape	Number of Params
Input Layer	Input	-	(224, 224, 3)	0
Residual Block 1	Convolutional	(7 × 7), stride 2	(112, 112, 64)	X (bergantung pada BiT-M)
Residual Block 2	Convolutional	(3 × 3), stride 2	(56, 56, 128)	X
Residual Block 3	Convolutional	(3 × 3), stride 2	(28, 28, 256)	X
Residual Block 4	Convolutional	(3 × 3), stride 2	(14, 14, 512)	X
Global Average Pooling	Pooling	-	(1, 1, 512)	0
Fully Connected	Dense	-	(N classes)	Y

   c. *Fine-tuning*: Model di-*fine-tune* pada dataset poster film dengan *learning rate* kecil untuk mengadaptasi bobot *pre-trained* ke domain spesifik tanpa kehilangan kemampuan generalisasi secara drastis.

2. **Model CLIP:**
   a. *Arsitektur*: Menggunakan model CLIP *pre-trained* yang menggabungkan *image encoder* (berbasis *Vision Transformer* - ViT atau ResNet) dan *text encoder* (berbasis Transformer). Tiga varian *image encoder* dieksplorasi: ViT-B/16, ViT-L/14, dan RN50x16. *Text encoder* digunakan untuk memproses deskripsi genre dan plot film.
   b. *Mekanisme*: CLIP bekerja dengan memproyeksikan *embedding* gambar (dari poster) dan *embedding* teks (dari deskripsi genre/plot) ke dalam ruang fitur bersama. Klasifikasi dilakukan dengan menghitung kesamaan (misalnya, *cosine similarity*) antara *embedding* gambar dan *embedding* teks untuk setiap genre target. Pendekatan ini memungkinkan klasifikasi *zero-shot* atau *few-shot*, namun dalam penelitian ini, model juga di-*fine-tune*.
   c. *Implementasi*: Gambar poster dan teks (deskripsi genre dan plot) diproses melalui *encoder* masing-masing. Untuk pendekatan *multi-modal* (poster + plot), *embedding* dari kedua modalitas dapat digabungkan sebelum klasifikasi akhir. Pelatihan menggunakan *contrastive loss* atau diadaptasi dengan *loss function* seperti *Binary Cross-Entropy* pada tahap *fine-tuning*.

E. Desain Eksperimen dan Hyperparameter

Eksperimen sistematis dilakukan untuk menemukan konfigurasi optimal bagi kedua model. Variasi *hyperparameter* yang diuji meliputi:

a. **Batch Size**: 16, 32, 64. Ukuran *batch* yang lebih kecil seringkali dikaitkan dengan generalisasi yang lebih baik (*flat minima*), namun perlu diseimbangkan dengan efisiensi komputasi.
b. **Learning Rate**: 0.001, 0.0001, 0.00001. Rentang ini dieksplorasi untuk menemukan laju pembelajaran yang stabil dan efektif.
c. **Optimizer**:
   i. Untuk CNN (BiT): Adam, SGD, RMSprop.
   ii. Untuk CLIP: AdamW (umum digunakan untuk model Transformer).

Syntax:
Journal of Software Engineering, Computer Science and Information Technology
p-ISSN: 2776-7027, e-ISSN: 2723-0538
Volume: 6, Nomor: 1, Juni 2025

d. Arsitektur CLIP: ViT-B/16, ViT-L/14, RN50x16.

Total kombinasi eksperimen adalah 27 untuk CNN (3 *batch sizes* x 3 *learning rates* x 3 *optimizers*) dan 27 untuk CLIP (3 *batch sizes* x 3 *learning rates* x 3 *architectures*). Pelatihan dilakukan selama maksimal 50 epoch (atau hingga *early stopping* terpicu).

Callback yang digunakan selama pelatihan:

a. EarlyStopping: Menghentikan pelatihan jika metrik pada *validation set* (misal, val_loss) tidak membaik selama beberapa epoch (*patience* = 5 atau 10).
b. ReduceLROnPlateau: Mengurangi *learning rate* (faktor 0.1 atau 0.2) jika metrik validasi stagnan (*patience* = 3 atau 5).
c. ModelCheckpoint: Menyimpan bobot model dengan performa terbaik pada *validation set*.

F. Metrik Evaluasi

Performa model dievaluasi menggunakan metrik standar untuk *multi-label classification* pada *testing set*:

a. **Akurasi**: Persentase prediksi yang benar secara keseluruhan (mungkin kurang informatif untuk *multi-label*).
b. **Hamming Loss**: Rata-rata proporsi label yang salah prediksi per sampel. Nilai lebih rendah lebih baik (rentang 0-1).
c. **F1-Score (Micro, Macro, Weighted)**: Rata-rata harmonik dari *precision* dan *recall*. *Micro* menghitung metrik secara global, *Macro* menghitung per label lalu dirata-rata (tanpa bobot), *Weighted* sama seperti *Macro* tetapi diboboti oleh jumlah sampel per label. Digunakan untuk mengukur keseimbangan performa.
d. **ROC AUC Score (Macro, Weighted)**: Area di bawah kurva ROC, mengukur kemampuan model membedakan antar kelas. Nilai lebih tinggi lebih baik (rentang 0-1).
e. **Classification Report**: Memberikan *precision, recall, f1-score*, dan *support* untuk setiap genre target secara individual.
f. **Loss Function (Binary Cross-Entropy)**: Nilai *loss* pada *training* dan *validation set* dipantau selama pelatihan untuk melihat konvergensi model.

Metrik-metrik ini memberikan evaluasi komprehensif tentang kemampuan model dalam mengklasifikasikan genre film secara akurat, menangani ketidakseimbangan kelas, dan mengukur kesalahan spesifik pada tugas *multi-label*.

III. HASIL DAN PEMBAHASAN

Tahap ini menyajikan hasil eksperimen dari pelatihan dan pengujian model CNN (BiT-ResNet50) dan CLIP (dengan variasi arsitektur ViT-B/16, ViT-L/14, RN50x16) pada dataset 15.150 poster film dengan 10 genre target. Analisis difokuskan pada perbandingan performa kedua pendekatan berdasarkan metrik evaluasi yang telah ditentukan.

A. Hasil Eksperimen Model CLIP

Model CLIP diuji dengan 27 kombinasi *hyperparameter* (3 *batch sizes* x 3 *learning rates* x 3 *architectures*) menggunakan *optimizer* AdamW. Ringkasan hasil terbaik untuk setiap arsitektur (berdasarkan kombinasi *Hamming Loss* terendah dan ROC AUC tertinggi pada *test set*) disajikan pada tabel 3 dan tabel 4.

Tabel 3
Eksperimen CLIP bagian 1

No	Model	Batch Size	Learning Rate	Test Accuracy	Hamming Loss
1	ViT-B/16	16	0,001	0,8279	0,1721
2	ViT-L/14	16	0,001	0,8304	0,1696
3	RN50x16	16	0,001	0,8235	0,1765
4	ViT-B/16	16	0,0001	0,8266	0,1734
5	ViT-L/14	16	0,0001	0,8321	0,1679
6	RN50x16	16	0,0001	0,8187	0,1813
7	ViT-B/16	16	0,00001	0,7525	0,2475
8	ViT-L/14	16	0,00001	0,7724	0,2276
9	RN50x16	16	0,00001	0,7464	0,2536
10	ViT-B/16	32	0,001	0,8282	0.1718
11	ViT-L/14	32	0,001	0,8315	0.1685
12	RN50x16	32	0,001	0,8241	0.1759
13	ViT-B/16	32	0,0001	0,8226	0.1774
14	ViT-L/14	32	0,0001	0,8322	0.1678
15	RN50x16	32	0,0001	0,8122	0.1878
16	ViT-B/16	32	0,00001	0,7503	0.2497
17	ViT-L/14	32	0,00001	0,7652	0.2348
18	RN50x16	32	0,00001	0,7486	0.2514
19	ViT-B/16	64	0,001	0,8261	0,1739
20	ViT-L/14	64	0,001	0,8326	0,1674
21	RN50x16	64	0,001	0,8243	0,1757
22	ViT-B/16	64	0,0001	0,8212	0,1788
23	ViT-L/14	64	0,0001	0,8083	0,1917
24	RN50x16	64	0,0001	0,7631	0,2369
25	ViT-B/16	64	0,00001	0,7501	0,2499
26	ViT-L/14	64	0,00001	0,7546	0,2454
27	RN50x16	64	0,00001	0,7432	0,2568

Tabel 4
Eksperimen CLIP bagian 2

No	Model	ROC AUC	Macro Average Precision	Last Epoch	Last Learning Rate
1	ViT-B/16	0,88	0,6859	6	0,001
2	ViT-L/14	0,89	0,7081	6	0,0005
3	RN50x16	0,88	0,6816	9	0,001
4	ViT-B/16	0,88	0,6840	15	0,0001
5	ViT-L/14	0,89	0,7098	10	0,0001
6	RN50x16	0,88	0,6708	21	0,0001
7	ViT-B/16	0,76	0,3986	8	0,000005

Syntax:
Journal of Software Engineering, Computer Science and Information Technology
p-ISSN: 2776-7027, e-ISSN: 2723-0538
Volume: 6, Nomor: 1, Juni 2025

8	ViT-L/14	0,8	0,5044	8	0,000005
9	RN50x16	0,75	0,3581	6	0,00001
10	ViT-B/16	0.88	0.6860	7	0,001
11	ViT-L/14	0.89	0.7092	6	0,0005
12	RN50x16	0.88	0.6827	10	0.001
13	ViT-B/16	0.88	0.6785	14	0,0001
14	ViT-L/14	0.89	0.7095	15	0,0001
15	RN50x16	0.87	0.6649	21	0,0001
16	ViT-B/16	0.76	0.3960	8	0,000005
17	ViT-L/14	0.78	0.4676	8	0,00001
18	RN50x16	0.74	0.3528	6	0,00001
19	ViT-B/16	0,88	0,6859	7	0,001
20	ViT-L/14	0,89	0,7097	6	0,001
21	RN50x16	0,88	0,6800	11	0,001
22	ViT-B/16	0,88	0,6758	19	0,0001
23	ViT-L/14	0,87	0,6506	8	0,00005
24	RN50x16	0,78	0,5107	7	0,00005
25	ViT-B/16	0,75	0,3681	6	0,00001
26	ViT-L/14	0,77	0,4245	8	0,00001
27	RN50x16	0,74	0,3327	6	0,00001

Performa Model CLIP yang ditunjukkan pada gambar 2 menunjukkan performa yang sangat baik secara keseluruhan, dengan akurasi mencapai 83,2% pada konfigurasi terbaik.

```
=== Classification Report ===
              precision    recall  f1-score   support

       Drama       0.77      0.92      0.84      1595
      Comedy       0.82      0.71      0.76      1094
      Action       0.76      0.68      0.72       887
       Crime       0.70      0.50      0.59       757
    Thriller       0.52      0.26      0.34       609
     Romance       0.72      0.55      0.62       621
   Adventure       0.77      0.34      0.47       429
      Horror       0.74      0.51      0.61       342
      Family       0.62      0.25      0.36       268
 Documentary       0.88      0.25      0.39       167

   micro avg       0.75      0.62      0.68      6769
   macro avg       0.73      0.50      0.57      6769
weighted avg       0.74      0.62      0.65      6769
 samples avg       0.76      0.62      0.66      6769
```

Gambar. 2 Hasil Test CLIP

Berdasarkan gambar 2, tabel 3 dan tabel 4 beberapa temuan utama dari hasil ekperimen CLIP sebagai berikut:

1. **CLIP (ViT-L/14)**: Menunjukkan performa terbaik secara keseluruhan. Konfigurasi optimal dicapai dengan *batch size* 32 dan *learning rate* 0.0001, menghasilkan:
   a. Akurasi (subset accuracy): 83,2%
   b. Hamming Loss: 0.1678
   c. ROC AUC (Macro): 0.89
   d. F1-Score (Weighted Avg): 0.65
   e. F1-Score (Micro Avg): 0.68
2. **CLIP (ViT-B/16)**: Menunjukkan performa baik, sedikit di bawah ViT-L/14. Hasil terbaik juga umumnya pada *learning rate* 0.0001 dan *batch size* 32.
3. **CLIP (RN50x16)**: Cenderung memiliki performa sedikit di bawah varian ViT, meskipun masih kompetitif.

Temuan kunci dari eksperimen CLIP:

1. Arsitektur **ViT-L/14** secara konsisten unggul, menunjukkan kapasitas model yang lebih besar bermanfaat untuk tugas ini.
2. *Learning rate* **0.0001** memberikan hasil paling optimal di semua arsitektur.
3. *Batch size* **32** menawarkan keseimbangan terbaik antara akurasi dan stabilitas pelatihan.

Analisis performa per genre untuk model CLIP terbaik (ViT-L/14, bs=32, lr=0.0001), berdasarkan Gambar 2 :

a. **Performa Tinggi**: Genre Drama (F1=0.84), Comedy (F1=0.76), dan Action (F1=0.72) diklasifikasikan dengan sangat baik.
b. **Performa Rendah**: Genre Thriller (F1=0.34), Family (F1=0.36), dan Documentary (F1=0.39) menunjukkan kesulitan signifikan. Rendahnya *recall* pada Documentary (0.25) dan Family (0.25) menunjukkan model sering melewatkan genre ini.
c. **Kesimpulan Performa CLIP**: Model sangat kuat untuk genre umum tetapi kesulitan pada genre yang lebih niche atau ambigu secara visual karena keterbatasan data atau fitur visual yang kurang unik.

B. Hasil Eksperimen CNN (BiT)

Model CNN (BiT-ResNet50) diuji dengan 27 kombinasi *hyperparameter* (3 *batch sizes* x 3 *learning rates* x 3 *optimizers*) menggunakan *optimizer* Adam,SGD, dan RMSprop. Ringkasan hasil terbaik untuk setiap arsitektur (berdasarkan kombinasi *Hamming Loss* terendah dan ROC AUC tertinggi pada *test set*) disajikan pada tabel 5 dan tabel 6.

Tabel 5
Eksperimen CNN bagian 1

No	Batch Size	Learning Rate	Optimizer	Test Loss	Test Accuracy	Test AUC
1	16	0,001	Adam	0,598	0,687	0,696
2	16	0,001	SGD	0,490	0,779	0,820
3	16	0,001	RMSprop	0,548	0,723	0,737
4	16	0,0001	Adam	0,506	0,771	0,806
5	16	0,0001	SGD	0,528	0,754	0,786
6	16	0,0001	RMSprop	0,479	0,778	0,811
7	16	0,00001	Adam	0,513	0,777	0,799
8	16	0,00001	SGD	0,581	0,713	0,764
9	16	0,00001	RMSprop	0,483	0,777	0,804
10	32	0,001	Adam	1,083	0,655	0,649
11	32	0,001	SGD	0,531	0,768	0,796
12	32	0,001	RMSprop	0,512	0,759	0,768
13	32	0,0001	Adam	0,506	0,770	0,807

Syntax:
Journal of Software Engineering, Computer Science and Information Technology
p-ISSN: 2776-7027, e-ISSN: 2723-0538
Volume: 6, Nomor: 1, Juni 2025

14	32	0,0001	SGD	0,504	0,778	0,794
15	32	0,0001	RMSprop	0,476	0,779	0,807
16	32	0,00001	Adam	0,563	0,768	0,789
17	32	0,00001	SGD	0,620	0,667	0,734
18	32	0,00001	RMSprop	0,500	0,779	0,802
19	64	0,001	Adam	0,589	0,724	0,715
20	64	0,001	SGD	0,570	0,733	0,779
21	64	0,001	RMSprop	0,506	0,754	0,771
22	64	0,0001	Adam	0,619	0,756	0,775
23	64	0,0001	SGD	0,562	0,732	0,787
24	64	0,0001	RMSprop	0,506	0,764	0,784
25	64	0,00001	Adam	0,662	0,681	0,772
26	64	0,00001	SGD	0,622	0,665	0,733
27	64	0,00001	RMSprop	0,526	0,758	0,773

Tabel 6
Eksperimen CNN bagian 2

No	Batch Size	Model AUC	Hamming Loss	Macro Average Precision	Last Epoch	Last Learning Rate
1	16	0,696	0,313	0,43	21	0,000125
2	16	0,820	0,221	0,56	15	0,00025
3	16	0,737	0,277	0,42	21	0,000125
4	16	0,806	0,229	0,52	15	0,000025
5	16	0,786	0,246	0,55	29	0,000012
6	16	0,811	0,222	0,50	19	0,000012
7	16	0,799	0,223	0,52	28	0,000002
8	16	0,764	0,287	0,54	58	0,000001
9	16	0,804	0,223	0,52	24	0,000002
10	32	0,649	0,345	0,40	25	0,000125
11	32	0,796	0,232	0,54	19	0,000125
12	32	0,768	0,241	0,40	16	0,00025
13	32	0,807	0,230	0,52	15	0,000025
14	32	0,794	0,222	0,55	33	0,000025
15	32	0,807	0,221	0,49	23	0,000025
16	32	0,789	0,232	0,53	43	0,000001
17	32	0,734	0,333	0,51	33	0,000001
18	32	0,802	0,221	0,53	34	0,000002
19	64	0,715	0,276	0,33	12	0,00025
20	64	0,779	0,267	0,55	18	0,00025
21	64	0,771	0,246	0,41	19	0,00025
22	64	0,775	0,244	0,52	16	0,000025
23	64	0,786	0,268	0,54	22	0,000012
24	64	0,784	0,236	0,53	15	0,000025
25	64	0,772	0,319	0,56	13	0,000002
26	64	0,733	0,335	0,48	23	0,000002
27	64	0,773	0,242	0,52	59	0,000001

Performa Model CNN (BiT) yang ditunjukkan pada gambar 3 menunjukkan hasil yang lebih bervariasi dibandingkan CLIP.

```
--- Classification Report (Test Set) ---
              precision    recall  f1-score   support

       Drama       0.77      0.86      0.81      1595
      Comedy       0.74      0.68      0.71      1094
      Action       0.72      0.51      0.60       887
       Crime       0.59      0.44      0.50       757
    Thriller       0.46      0.60      0.52       609
     Romance       0.67      0.40      0.50       621
   Adventure       1.00      0.00      0.01       429
      Horror       0.49      0.49      0.49       342
      Family       0.49      0.19      0.27       268
 Documentary       0.35      0.29      0.32       167

   micro avg       0.66      0.56      0.61      6769
   macro avg       0.63      0.45      0.47      6769
weighted avg       0.68      0.56      0.58      6769
 samples avg       0.67      0.56      0.59      6769
```

Gambar. 3 Hasil test CNN

Berdasarkan gambar 3, tabel 5 dan tabel 6 beberapa temuan utama dari hasil eksperimen CNN (BiT) sebagai berikut:

1. **Konfigurasi Terbaik**: Dicapai dengan *optimizer* **SGD**, *learning rate* **0.001**, dan *batch size* **16**, menghasilkan:
   a. Akurasi (subset accuracy): 77,9%
   b. Hamming Loss: 0,221 (nilai lebih tinggi dari CLIP).
   c. ROC AUC (Macro): 0,82 (nilai lebih rendah dari CLIP)
   d. F1-Score (Weighted Avg): 0.58
   e. F1-Score (Micro Avg): 0.61

Temuan kunci dari eksperimen CNN (BiT):

1. *Optimizer* **SGD** memberikan performa terbaik dibandingkan Adam dan RMSprop untuk arsitektur BiT pada dataset ini.
2. *Learning rate* **0.001** (relatif tinggi untuk *fine-tuning*) menghasilkan hasil paling stabil.
3. *Batch size* **16** memberikan hasil terbaik, meskipun terindikasi adanya sedikit *overfitting* pada epoch akhir.

Analisis performa per genre untuk model CNN terbaik (SGD, lr=0.001, bs=16), berdasarkan Gambar 3:

a. **Performa Tinggi**: Genre Drama (F1=0.81) dan Comedy (F1=0.71) masih menjadi yang terkuat.
b. **Performa Stabil**: Action (F1=0.60).
c. **Performa Rendah**: Genre Adventure (F1=0.00 karena *recall*=0.00), Family (F1=0.27), dan Documentary (F1=0.32) menunjukkan kesulitan yang lebih parah dibandingkan CLIP. Thriller (F1=0.52) juga relatif rendah.
d. **Kesimpulan Performa CNN (BiT)**: Model menunjukkan performa yang layak untuk genre umum, namun secara signifikan lebih rendah daripada CLIP, terutama pada genre yang lebih

Syntax:
Journal of Software Engineering, Computer Science and Information Technology
p-ISSN: 2776-7027, e-ISSN: 2723-0538
Volume: 6, Nomor: 1, Juni 2025

sulit. Kegagalan total pada Adventure (*recall*=0) menjadi perhatian khusus.

C. Perbandingan Performa CLIP dan CNN (BiT)

Perbandingan langsung antara model terbaik dari kedua pendekatan menyoroti keunggulan signifikan CLIP:

1. **Metrik Keseluruhan**: CLIP (ViT-L/14) secara konsisten mengungguli CNN (BiT) di hampir semua metrik utama: *Hamming Loss* lebih rendah (0.1678 vs >0.221), ROC AUC lebih tinggi (0.89 vs <0.82), dan F1-Score *Weighted Average* lebih tinggi (0.65 vs 0.58).
2. **Performa Per Genre**: Meskipun kedua model kuat pada Drama dan Comedy, CLIP menunjukkan keunggulan pada Action dan secara signifikan lebih baik, meskipun masih rendah pada genre seperti Thriller, Family, dan Documentary dibandingkan CNN. CNN bahkan gagal total pada Adventure.
3. **Kemampuan Multi-Modal**: Keunggulan CLIP berasal dari kemampuan untuk memanfaatkan informasi teks (deskripsi genre dan plot) bersama dengan informasi visual poster. Representasi *multi-modal* ini memungkinkan pemahaman konteks yang lebih kaya dibandingkan CNN yang hanya mengandalkan fitur visual. CNN mungkin kesulitan membedakan genre dengan estetika visual yang tumpang tindih atau ambigu tanpa konteks tambahan.
4. **Integrasi Plot (BERT)**: Penelitian ini juga mencatat bahwa penambahan analisis plot menggunakan BERT (sebagai bagian dari alur kerja CLIP) meningkatkan akurasi sekitar 5% dibandingkan klasifikasi hanya berbasis poster. Ini semakin memperkuat argumen untuk pendekatan *multi-modal*.

D. Analisis Prediksi pada Data Baru

Pada gambar 4 adalah film "Captain America: Brave New World" (rilis 2025, tidak ada di dataset) yang akan diprediksi sebagai dataset untuk menguji kedua model terbaik.

Gambar. 4 Captain America

Hasil prediksi disajikan pada gambar 5 dengan model CNN terbaik dan gambar 6 dengan model CLIP terbaik.

```
Hasil Prediksi (diurutkan dari tertinggi ke terendah):
Action: 0.8741
Thriller: 0.6469
Drama: 0.6196
Comedy: 0.4969
Adventure: 0.3961
Crime: 0.1738
Romance: 0.1304
Documentary: 0.1230
Family: 0.0410
Horror: 0.0245
```

Gambar. 5 Hasil CNN

```
Predicted Genres (sorted by probability):
Action: 0.87
Drama: 0.52
Crime: 0.51
Adventure: 0.49
Comedy: 0.36
Thriller: 0.18
Romance: 0.03
Family: 0.03
Horror: 0.01
Documentary: 0.00
```

Gambar. 6 Hasil CLIP

Hasilnya Gambar 5 dan gambar 6 menunjukkan prediksi probabilitas yang **identik** antara CNN dan CLIP yaitu Action sebesar 0.87 yang merupakah hasil prediksi tertinggi. Kesamaan hasil ini menunjukkan bahwa untuk poster dengan sinyal visual genre Action yang sangat kuat, kedua model konvergen pada interpretasi yang sama. Namun, ini tidak meniadakan keunggulan CLIP secara keseluruhan yang terbukti pada evaluasi *test set* yang lebih luas. Perbedaan mungkin akan lebih terlihat pada contoh film dengan

visual yang ambigu atau memerlukan pemahaman konteks teks yang lebih dalam.

Secara keseluruhan, hasil eksperimen dengan jelas menunjukkan bahwa model CLIP, terutama dengan arsitektur ViT-L/14 dan memanfaatkan pendekatan *multi-modal* dengan analisis plot, secara signifikan lebih unggul daripada model CNN (BiT-ResNet50) untuk tugas klasifikasi genre film *multi-label* berbasis poster.

KESIMPULAN DAN SARAN

A. Kesimpulan

Penelitian ini berhasil mengimplementasikan dan mengevaluasi dua pendekatan *deep learning*, yaitu *Convolutional Neural Network* (CNN) menggunakan *Transfer Learning* dengan BiT-ResNet50 dan model *multi-modal Contrastive Language-Image Pretraining* (CLIP) dengan berbagai arsitektur (*Vision Transformer* dan ResNet), untuk tugas klasifikasi genre film *multi-label* berdasarkan analisis poster dan plot film. Berdasarkan analisis hasil eksperimen pada dataset 15.150 film dari IMDb dan OMDb, beberapa kesimpulan utama dapat ditarik:

1. Model CLIP menunjukkan performa yang secara signifikan lebih unggul dibandingkan model CNN (BiT). Konfigurasi CLIP terbaik (menggunakan arsitektur ViT-L/14, *batch size* 32, dan *learning rate* 0.0001) mencapai F1-Score (*Weighted Average*) sebesar 0.65 dan *Hamming Loss* 0.1678 pada *test set*. Sebagai perbandingan, konfigurasi CNN (BiT) terbaik (menggunakan *optimizer* SGD, *batch size* 16, dan *learning rate* 0.001) hanya mencapai F1-Score (*Weighted Average*) sebesar 0.58.
2. Keunggulan CLIP terutama disebabkan oleh kemampuannya dalam memahami hubungan semantik antara informasi visual (poster) dan tekstual (deskripsi genre dan plot) melalui representasi *multi-modal*. Pendekatan ini memungkinkan CLIP menangkap konteks yang lebih kaya dibandingkan CNN yang hanya mengandalkan fitur visual, sehingga lebih efektif dalam membedakan genre, terutama untuk kasus *multi-label* di mana sebuah film dapat memiliki beberapa genre.
3. Integrasi analisis plot film menggunakan model BERT sebagai bagian dari alur kerja CLIP terbukti meningkatkan akurasi klasifikasi sekitar 5% dibandingkan dengan metode yang hanya berbasis analisis poster. Hal ini mengkonfirmasi manfaat signifikan dari pendekatan *multi-modal learning* untuk tugas ini.
4. Meskipun kedua model menunjukkan performa terbaik pada genre umum seperti Drama, Comedy, dan Action, keduanya masih menghadapi kesulitan pada genre yang lebih niche atau memiliki representasi visual yang ambigu, seperti Thriller, Family, dan Documentary. Performa CLIP pada genre-genre ini, meskipun rendah, masih lebih baik dibandingkan CNN.
5. Pemilihan *hyperparameter* memiliki dampak signifikan. Untuk CLIP, *learning rate* 0.0001 dan *batch size* 32 terbukti optimal. Untuk CNN (BiT), *optimizer* SGD dengan *learning rate* 0.001 dan *batch size* 16 memberikan hasil terbaik.

Secara keseluruhan, penelitian ini menegaskan bahwa model *Vision-Language* seperti CLIP menawarkan solusi yang lebih efektif dan akurat untuk klasifikasi genre film *multi-label* dibandingkan pendekatan CNN konvensional berbasis visual murni. Kemampuan untuk mengintegrasikan dan memahami informasi dari berbagai modalitas menjadi kunci keberhasilan dalam menangani kompleksitas tugas ini.

B. Saran

Berdasarkan hasil penelitian yang telah dilakukan, terdapat beberapa area yang dapat dikembangkan untuk meningkatkan performa sistem klasifikasi genre film multi-label:

1. Implementasi fine-tuning pada model CLIP menggunakan dataset film yang lebih luas dan beragam berpotensi meningkatkan akurasi klasifikasi secara signifikan. Fine-tuning yang berfokus pada karakteristik visual spesifik poster film dapat mengoptimalkan kemampuan model dalam mengenali elemen-elemen distintif setiap genre, seperti palet warna, komposisi visual, dan objek-objek khas yang menjadi penanda genre tertentu.
2. Eksplorasi penggunaan judul film sebagai fitur tambahan melalui pendekatan multimodal dapat memberikan konteks semantik yang lebih kaya untuk proses klasifikasi. Integrasi Natural Language Processing (NLP) pada judul film dengan analisis visual poster dan plot dapat menciptakan representasi yang lebih komprehensif, mengingat judul film seringkali mengandung kata kunci atau frasa yang mengindikasikan genre spesifik.

Pengembangan lebih lanjut dari penelitian ini diharapkan dapat memberikan kontribusi signifikan bagi industri perfilman dan platform streaming dalam mengotomatisasi proses klasifikasi konten, sekaligus meningkatkan efektivitas sistem rekomendasi berbasis genre yang lebih akurat dan personal.

REFERENSI

[1] S. Pooranalingam, "Film Poster Design: Understanding Film Poster Designs and the Compositional Similarities within specific genres," *Spectrum*, no. 12, Jan. 2024, doi: 10.29173/spectrum216.

[2] W. T. Chu and H. J. Guo, "Movie genre classification based on poster images with deep neural networks," in *MUSA2 2017 - Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes, co-located with MM 2017*, Association for Computing Machinery, Inc, Oct. 2017, pp. 39–45. doi: 10.1145/3132515.3132516.

[3] N. Hossain, M. M. Ahamad, S. Aktar, and M. A. Moni, "Movie Genre Classification with Deep Neural Network using Poster Images," in *2021 International Conference on Information and Communication Technology for Sustainable Development, ICICT4SD 2021 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., Feb. 2021, pp. 195–199. doi: 10.1109/ICICT4SD50815.2021.9396778.

[4] G. Barney and K. Kaya, "Predicting Genre from Movie Posters," CS229 Machine Learning Project Report, Stanford University, 2019. [Online]. Available: https://cs229.stanford.edu/proj2019spr/report/9.pdf

[5] J. A. Wi, S. Jang, and Y. Kim, "Poster-Based Multiple Movie Genre Classification Using Inter-Channel Features," *IEEE Access*, vol. 8, pp. 66615–66624, 2020, doi: 10.1109/ACCESS.2020.2986055.

[6] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," Feb. 2021. [Online]. Available: http://arxiv.org/abs/2103.00020.

[7] S. Shen et al., "How Much Can CLIP Benefit Vision-and-Language Tasks?," Jul. 2021. [Online]. Available: http://arxiv.org/abs/2107.06383.

[8] I. H. Witten, E. Frank, and M. A. Hall, Data Mining Practical Machine Learning Tools and Technique. Burlington: Morgan Kaufmann Publisher, 2011.