

BAB V PENUTUP

5.1 Kesimpulan

Berdasarkan hasil analisis perbandingan kinerja secara *head-to-head* antara arsitektur 3D CNN, Video Vision Transformer (ViViT), dan 3D ConvNeXt dengan batasan dataset berskala menengah, maka dapat ditarik kesimpulan yang menjawab rumusan masalah penelitian sebagai berikut:

1. Perbandingan Kinerja Akurasi: Berdasarkan hasil pengujian klasifikasi video *Deepfake*, arsitektur 3D CNN murni menunjukkan kinerja yang paling unggul dengan tingkat akurasi mencapai 80,00%, Recall 90,00%, dan nilai AUC-ROC 89,81%. Sebaliknya, arsitektur modern *State-of-the-Art* (SOTA) seperti ViViT dan 3D ConvNeXt hanya mampu mencatatkan akurasi di kisaran 51% (51.00% untuk ViViT dan 49.67% untuk 3D ConvNeXt), yang praktis setara dengan probabilitas tebakan acak.
2. Superioritas Deteksi Artefak Spasio-Temporal: Di antara ketiga arsitektur tersebut, 3D CNN merupakan model yang paling unggul dalam mendeteksi artefak spasio-temporal yang halus pada video *Deepfake*. Keunggulan ini terjadi karena 3D CNN memiliki *inductive bias* spasio-temporal bawaan yang mampu mengenali pola anomali lokal (seperti kesalahan blending wajah) secara cepat. Sementara itu, ViViT dan 3D ConvNeXt gagal menangkap ciri khas tersebut karena mengalami kekurangan data (*data starvation*) dan *mode collapse*; ketiadaan *inductive bias* pada arsitektur berbasis *Transformer* mengharuskan mereka untuk belajar dari dataset berskala sangat masif untuk bisa menemukan fitur anomali.
3. Kelebihan dan Kekurangan Model pada Skala Data Kecil: Dalam menyeimbangkan akurasi deteksi dan efisiensi waktu komputasi (*inference time*) pada dataset berskala kecil, ditemukan karakteristik sebagai berikut:
 - 3D CNN: Memiliki kelebihan utama pada keseimbangan yang sangat proporsional antara akurasi yang andal (80,00%) dan efisiensi komputasi

yang ringan (waktu inferensi 0,0531 detik per video), menjadikannya model yang paling praktis.

- **ViViT:** Memiliki kelebihan berupa kecepatan inferensi tertinggi saat *batch processing* di GPU (911.10 FPS) berkat paralelisasi mekanisme Self-Attention. Namun, arsitektur ini memiliki kelemahan fatal berupa rendahnya daya tangkap fitur (*Recall* 3.33%) jika tidak didukung oleh jutaan data latih.
- **3D ConyNeXt:** Menawarkan rancang bangun parameter yang relatif ringan layaknya model hibrida modern, namun memiliki kelemahan berupa kerentanan yang tinggi terhadap *over-prediction* (bias menebak satu kelas secara terus-menerus) saat dihadapkan pada dimensi temporal video.

5.2 Saran

Berdasarkan proses evaluasi dan analisis kesalahan (*error analysis*) selama penelitian, terdapat beberapa saran terfokus untuk pengembangan penelitian selanjutnya:

1. **Peningkatan Skala Dataset:** Disarankan menggunakan dataset berskala masif (jutaan *frame*) dan komputasi tinggi untuk menguji potensi maksimal model berbasis *Transformer* (seperti ViViT) guna menghindari masalah *data starvation*.
2. **Optimasi Algoritma Pendeteksi Wajah:** Disarankan menggunakan algoritma pelacakan wajah (*face alignment*) yang lebih dinamis untuk menangani variasi pose ekstrem (rotasi wajah 3D) agar pemotongan bounding box lebih akurat dan tidak menghilangkan fitur krusial.
3. **Augmentasi Data Ekstrem:** Perlu diterapkan metode augmentasi data yang lebih agresif pada tahap pra-pemrosesan, khususnya pada variasi pencahayaan gelap dan kompresi resolusi, untuk meningkatkan ketahanan model terhadap kebingungan (False Alarm).
4. **Eksplorasi Modifikasi Hibrida:** Penelitian ke depan dapat mengeksplorasi penggabungan metode hibrida (CNN-Transformer) yang disesuaikan khusus

untuk arsitektur ringan (*lightweight*) guna mengatasi batasan komputasi tanpa mengorbankan akurasi klasifikasi.

