BAB V
PENUTUP

5.1 Kesimpulan

Berdasarkan hasil penelitian dan pengujian yang telah dilakukan, dapat disimpulkan bahwa arsitektur *Deep Learning* gabungan dari *Convolutional Neural Network* (CNN), *Bidirectional LSTM* (Bi-LSTM), dan *Transformer* telah berhasil dirancang dan diimplementasikan untuk mengolah fitur visual spasial dan temporal pada data suara. Model hibrida ini terbukti sangat andal dalam mengatasi tantangan keragaman data suara multi-aksen global untuk membedakan usia dan *gender*, di mana hasil pengujian menunjukkan capaian akurasi sebesar 91% yang didukung oleh nilai evaluasi *F1-Score* yang stabil dan konsisten. Capaian ini sekaligus menjawab tujuan penelitian terkait perbandingan efektivitas, di mana model gabungan terbukti memiliki kinerja yang secara signifikan lebih unggul dibandingkan jika hanya menggunakan model arsitektur tunggal pada kondisi *dataset* yang sama. Hal ini terlihat jelas dari perbandingan hasil akurasinya yang berhasil melampaui model CNN murni yang mendapatkan 84% serta model Bi-LSTM murni yang mendapatkan 74%. Dengan demikian, strategi menyatukan kemampuan membaca pola gambar spektral (fitur spasial), dinamika urutan waktu (temporal), dan pemusatan perhatian (*attention*) terbukti menjadi metode yang efektif untuk menyelesaikan kompleksitas klasifikasi karakteristik suara manusia.

5.2 Saran

Berdasarkan hasil penelitian dan kendala yang dihadapi selama proses pengembangan sistem, terdapat beberapa saran yang dapat diberikan untuk pengembangan penelitian selanjutnya agar menjadi lebih baik, antara lain:

1. Disarankan agar penelitian selanjutnya dapat menggunakan atau mengumpulkan dataset yang memiliki distribusi kelas lebih seimbang sejak awal, sehingga tidak mengalami masalah minimnya data pada kelas minoritas seperti pada kategori remaja dan lansia. Selain itu, penting juga untuk memastikan kualitas rekaman audio yang lebih jernih dan atribut

data yang lebih lengkap agar model dapat belajar dari sumber data yang lebih berkualitas.

2. Disebabkan model yang dibangun merupakan gabungan dari beberapa algoritma yang cukup rumit, proses pelatihannya menuntut kinerja komputer yang tinggi dan memakan waktu lama. Disarankan untuk menggunakan perangkat dengan kartu grafis (GPU) yang lebih mumpuni agar proses pembelajaran model bisa berjalan lebih cepat dan efisien tanpa kendala teknis yang berarti.
3. Disarankan untuk bereksperimen dengan model-model *Deep Learning* yang lebih modern dan canggih, seperti *EfficientNet* atau model yang memang dirancang khusus untuk data suara seperti *Wav2Vec 2.0*. Selain itu, dapat juga dicoba metode penggabungan beberapa model (*Ensemble Learning*) agar bisa saling melengkapi kekurangan masing-masing model, sehingga hasil prediksinya menjadi lebih akurat dibandingkan hanya mengandalkan satu model saja.
4. Pengembangan sistem pada tahap *deployment* saat ini masih berbasis web sederhana menggunakan Streamlit. Diharapkan pada pengembangan selanjutnya sistem dapat diimplementasikan ke dalam bentuk aplikasi *mobile* (Android/iOS) atau sistem *real-time* yang dapat memproses suara secara langsung dari mikrofon tanpa perlu mengunggah *file* audio terlebih dahulu.