BAB I
PENDAHULUAN

1.1 Latar Belakang

Seiring dengan pesatnya perkembangan teknologi informasi, cara manusia berinteraksi dengan mesin telah berubah secara signifikan, di mana sistem pintar saat ini diharapkan untuk tidak hanya mampu menjalankan perintah, tetapi juga dapat memahami siapa penggunanya agar dapat memberikan layanan yang lebih personal dan sesuai kebutuhan [1]. Kemampuan sistem untuk mengenali ciri dasar pengguna, seperti usia dan gender secara otomatis menjadi kunci utama dalam perkembangan ini. Dengan memahami demografi pengguna, sistem dapat menyesuaikan gaya komunikasi, konten, hingga alur layanan secara dinamis, sehingga membuka berbagai peluang baru untuk meningkatkan kualitas interaksi di berbagai sektor industri [1]. Pada akhirnya, tujuan dari pengembangan teknologi ini adalah untuk menciptakan interaksi antara manusia dan mesin yang terasa lebih alami dan cerdas, yang pada gilirannya dapat meningkatkan kepuasan dan loyalitas pengguna dalam jangka panjang.

Manfaat praktis dari teknologi ini sangatlah luas dan telah diterapkan dalam berbagai skenario di dunia nyata, mulai dari personalisasi konten hingga sistem keamanan akses. Dalam industri *e-commerce*, misalnya, asisten suara pada aplikasi belanja *online* dapat menganalisis suara pengguna untuk memperkirakan demografinya, kemudian menawarkan rekomendasi produk atau iklan yang lebih sesuai dengan kelompok usia dan gendernya, sebuah pendekatan yang terbukti dapat meningkatkan keterlibatan dan keputusan pembelian pengguna [2]. Contoh penerapan lainnya yang sangat relevan adalah pada sistem keamanan kendaraan, di mana deteksi usia dapat digunakan untuk membatasi akses, misalnya sistem dapat membedakan antara suara orang dewasa dan anak-anak untuk mencegah anak-anak menyalakan mobil tanpa pengawasan, sehingga meningkatkan aspek keselamatan [3].

Dalam upaya mencapai deteksi yang akurat, berbagai penelitian terdahulu

telah mengeksplorasi beragam metode klasifikasi. Namun, model sekuensial murni seperti *Recurrent Neural Network* (RNN) memiliki keterbatasan dalam menangani dependensi jangka panjang akibat masalah *vanishing gradient*, yang membuat kinerja model menurun seiring bertambahnya durasi rekaman [4]. Selain itu, data suara dalam bentuk *Mel-Spectrogram* memiliki karakteristik visual dua dimensi, sehingga RNN juga seringkali kurang optimal dalam mengekstrak fitur spasial atau tekstur visual tersebut secara langsung [5]. Di sisi lain, *Convolutional Neural Network* (CNN) terbukti efektif menangkap fitur lokal dari gambar spektogram tersebut, namun memiliki kelemahan dalam memodelkan urutan waktu yang kompleks jika berdiri sendiri [6]. Sementara itu, meskipun arsitektur *Transformer* sangat unggul dalam menangkap konteks global, kinerjanya terbukti lebih optimal jika disinergikan dengan pemodelan sekuensial seperti Bi-LSTM untuk menangkap dinamika temporal lokal secara mendetail, sebagaimana dibuktikan dalam studi klasifikasi suara yang menggunakan arsitektur hibrida serupa [7].

Berdasarkan tantangan tersebut, penelitian ini mengusulkan model gabungan (hibrida) yang menggabungkan arsitektur CNN, Bi-LSTM, dan *Transformer*. Dalam pendekatan ini, CNN difungsikan sebagai pengekstraksi fitur spasial untuk menangkap pola visual dari *Mel-Spectrogram*. Selanjutnya, *Bi-Directional LSTM* (Bi-LSTM) diterapkan untuk mempelajari ketergantungan urutan waktu secara dua arah guna mengatasi keterbatasan memori jangka pendek pada RNN konvensional. Terakhir, *Transformer* digunakan untuk memodelkan konteks global melalui mekanisme atensi. Penggabungan ketiga arsitektur ini saling melengkapi satu sama lain, di mana kelebihan masing-masing model bekerja secara bersamaan untuk mengatasi kerumitan data suara yang sangat beragam. Artinya, sistem tidak hanya mampu mengenali bentuk visual dari sinyal suara, tetapi juga memahami urutan waktunya serta bagian mana yang paling penting untuk diperhatikan [8]. Dengan demikian, model yang diusulkan diharapkan bisa menghasilkan pemahaman karakteristik suara yang utuh, mulai dari detail visualnya, alur waktunya, hingga gambaran besarnya, sehingga akurasi deteksi usia dan gender bisa meningkat secara signifikan dibandingkan jika hanya

menggunakan satu model saja.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana merancang dan mengimplementasikan arsitektur *hybrid* yang menggabungkan *Convolutional Neural Network* (CNN), *Bi-Directional LSTM* (Bi-LSTM), dan *Transformer* untuk mengolah fitur visual dan temporal pada data suara?
2. Bagaimana kinerja model hybrid CNN-BiLSTM-Transformer dalam mengatasi variabilitas data suara multi-aksen untuk klasifikasi usia dan gender berbasis Mel-Spectrogram?
3. Bagaimana perbandingan efektivitas model hybrid yang diusulkan terhadap arsitektur tunggal (seperti CNN murni dan Bi-LSTM) dalam menyelesaikan tugas klasifikasi tersebut?

1.3 Batasan Masalah

Agar penelitian ini lebih terarah dan pembahasannya tidak terlalu luas, maka penelitian ini dibatasi pada beberapa hal berikut:

1. Penelitian ini hanya menggunakan dataset *Mozilla Common Voice* yang didapatkan dari platform Kaggle. Hasil kinerja model nantinya akan sangat bergantung pada karakteristik dan kualitas data yang tersedia dalam dataset tersebut.
2. Pengolahan data suara dibatasi menggunakan metode *Mel-Spectrogram*. Penelitian ini tidak akan mencoba atau membandingkan teknik lain (seperti MFCC), karena fokus utamanya adalah memanfaatkan bentuk visual spektrum suara untuk diolah oleh model.
3. Model yang dibangun secara khusus menggabungkan tiga algoritma sekaligus, yaitu *CNN*, *Bi-LSTM*, dan *Transformer*. Variasi model lain seperti RNN biasa atau GRU tidak akan dibahas dalam penelitian ini.

4. Penelitian ini hanya berfokus untuk mengenali dua hal yaitu Gender (Pria dan Wanita) dan Kelompok Usia sesuai label yang ada di dataset. Hal-hal lain seperti aksen bicara, emosi, atau kondisi kesehatan penutur tidak termasuk dalam analisis.
5. Fokus utama penelitian adalah pada perancangan teknis sistem dan pengukuran akurasinya. Analisis mendalam mengenai isu etika atau dampak sosial dari teknologi ini tidak menjadi fokus utama pembahasan.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah di atas, tujuan dari penelitian ini adalah:

1. Merancang dan membangun arsitektur *Deep Learning* gabungan (hibrida) yang mengintegrasikan *Convolutional Neural Network* (CNN), *Bi-LSTM*, dan *Transformer* untuk melakukan klasifikasi usia dan gender dari data suara.
2. Mengukur seberapa baik kinerja model tersebut dalam mengenali usia dan gender menggunakan parameter evaluasi standar, yaitu akurasi, *precision*, *recall*, dan *f1-score*.
3. Menganalisis efektivitas model yang diusulkan dengan cara membandingkan hasil akurasinya melawan model yang hanya menggunakan satu arsitektur saja (seperti Bi-LSTM atau CNN murni).

1.5 Manfaat Penelitian

Adapun manfaat yang diharapkan dari penelitian ini adalah sebagai berikut:

1. Secara teoretis, penelitian ini diharapkan dapat memberikan kontribusi signifikan pada pengembangan ilmu kecerdasan buatan, khususnya dalam dalam bidang pemrosesan sinyal suara digital. Selain itu, kajian ini dapat berfungsi sebagai referensi teknis dan landasan ilmiah bagi peneliti selanjutnya yang ingin mengeksplorasi atau mengembangkan arsitektur *hybrid*, yang menggabungkan kekuatan *Convolutional Neural Network*

(CNN), *Bi-Directional LSTM* (Bi-LSTM), dan *Transformer*, untuk menyelesaikan berbagai permasalahan klasifikasi audio yang kompleks.

2. Secara praktis, solusi teknologi yang dihasilkan memiliki potensi penerapan yang luas di berbagai sektor industri. Dalam bidang *e-commerce*, model ini dapat memfasilitasi personalisasi layanan yang lebih cerdas dengan merekomendasikan produk berdasarkan prediksi demografi pelanggan. Di sektor keamanan, teknologi ini dapat memperkuat sistem akses pada kendaraan atau rumah pintar dengan membedakan hak akses berdasarkan karakteristik suara pengguna. Lebih jauh lagi, dalam bidang kriminologi dan forensik digital, sistem ini dapat dimanfaatkan sebagai alat bantu investigasi awal untuk memprediksi profil pelaku kejahatan melalui analisis bukti rekaman suara.

1.6 Sistematika Penulisan

Untuk memberikan gambaran yang jelas mengenai struktur dan isi dari laporan skripsi ini, maka penyusunannya diorganisasikan ke dalam lima bab utama dengan rincian sebagai berikut:

BAB I - PENDAHULUAN Bab ini menguraikan latar belakang masalah yang mendasari penelitian, perumusan masalah yang akan diselesaikan, batasan masalah untuk menjaga fokus pembahasan, tujuan yang ingin dicapai, serta manfaat yang diharapkan dari penelitian ini, baik secara teoretis maupun praktis.

BAB II - TINJAUAN PUSTAKA Bab ini memuat tinjauan terhadap penelitian-penelitian terdahulu yang relevan sebagai acuan komparasi. Selain itu, bab ini juga memaparkan landasan teori yang digunakan, meliputi konsep dasar pemrosesan sinyal suara (khususnya representasi *Mel-Spectrogram*), serta teori mendalam mengenai arsitektur *Deep Learning* yang diterapkan, yaitu *Convolutional Neural Network* (CNN), *Bi-Directional Long Short-Term Memory* (Bi-LSTM), dan *Transformer*.

BAB III - METODOLOGI PENELITIAN Bab ini menjelaskan tahapan metodologis sistematis yang dilakukan dalam penelitian. Pembahasan mencakup

objek penelitian, alat dan bahan (termasuk dataset), teknik pra-pemrosesan data, perancangan rinci arsitektur model hibrida (CNN-BiLSTM-Transformer), serta skenario pengujian dan metrik evaluasi yang digunakan untuk mengukur kinerja model.

BAB IV - HASIL DAN PEMBAHASAN Bab ini menyajikan hasil eksperimen yang telah dilakukan beserta analisisnya. Data hasil pengujian akan ditampilkan dalam bentuk tabel atau grafik untuk memudahkan pembacaan. Fokus utama bab ini adalah menganalisis kinerja model hibrida yang diusulkan dan membandingkannya dengan model arsitektur tunggal untuk menjawab rumusan masalah.

BAB V - PENUTUP Bab ini merupakan bagian akhir dari laporan skripsi yang berisi kesimpulan menyeluruh dari penelitian yang telah dilakukan, serta saran-saran konstruktif yang dapat digunakan sebagai acuan untuk pengembangan sistem atau penelitian lebih lanjut di masa mendatang.