Bab 1 Pendahuluan

Komunikasi merupakan kebutuhan dasar manusia, tetapi bagi individu dengan gangguan pendengaran, hal ini dapat menghadirkan tantangan yang signifikan terutama ketika penerjemah bahasa isyarat yang berkualifikasi tidak tersedia. Bahasa isyarat berfungsi sebagai media komunikasi utama dalam komunitas tuna rungu dan dicirikan oleh ekspresi visual-gestural yang kompleks. Namun, penerjemahan manual tetap menjadi hambatan bagi partisipasi penuh dalam domain-domain penting seperti pendidikan, layanan kesehatan, dan pekerjaan.

Meskipun penerjemah manusia dapat membantu menjembatani kesenjangan komunikasi antara orang yang dapat mendengar dan yang memiliki gangguan pendengaran, saat ini belum tersedia cukup penerjemah yang berkualifikasi. Situasi ini seringkali dipersulit oleh keterbatasan waktu dan anggaran. Terdapat teknik yang menjanjikan untuk menjembatani kesenjangan komunikasi ini dan mempermudah mereka yang memiliki gangguan pendengaran untuk berinteraksi dengan dunia luar [14]. Kemajuan dalam identifikasi bahasa isyarat otomatis menawarkan solusi potensial untuk masalah ini.

Pengenalan pola dan klasifikasi gambar, serta aplikasi dalam identifikasi gerakan tangan untuk bahasa isyarat, telah dipengaruhi secara signifikan oleh kemajuan terbaru dalam Pembelajaran Mendalam [3], [6]. Jaringan Saraf Konvolusional (CNN) adalah salah satu metode yang telah menunjukkan efektivitas yang sangat baik dalam menguraikan pola visual yang kompleks. CNN telah digunakan secara luas dalam penelitian yang berkaitan dengan pengenalan bahasa isyarat karena kemampuannya yang luar biasa untuk mengekstraksi karakteristik spasial dari gambar. Sebuah studi oleh Rao et al. [3] mengatakan bahwa CNN dapat memahami gerakan tangan dengan benar dalam banyak bahasa isyarat, seperti Bahasa Isyarat Amerika (ASL). Jaringan Memori Jangka Panjang (LSTM) dan teknik pemodelan urutan temporal lainnya juga digunakan untuk mempelajari bagaimana gerakan tangan berubah seiring waktu [7], [12]. Dengan menggabungkan CNN dan LSTM, model tersebut dapat merekam aspek ruang dan waktu dari data gerakan. Dikatakan oleh Das et al. bahwa metode campuran ini membuat klasifikasi lebih akurat daripada hanya menggunakan CNN atau LSTM. [4], [11].

Studi Prangon Das dkk. tahun 2020 merekomendasikan penggunaan CNN untuk mengenali gerakan tangan statis dalam ASL. Mereka melakukan ini dengan menggunakan satu set 1.815 foto dari 26 kelas alfabet. Model tersebut mengungguli berbagai teknik terkini dengan akurasi validasi sebesar 94,34%. Meskipun demikian, studi tersebut hanya mencakup gerakan statis dan set data kecil, yang mungkin tidak secara akurat mencerminkan daya tarik dan keragaman bahasa isyarat dalam kehidupan sehari-hari. Lebih lanjut, dinamika temporal yang krusial untuk mengidentifikasi ekspresi bahasa isyarat yang berkelanjutan atau dinamis tidak disertakan dalam model [4]. Sebuah studi berbeda oleh Kamruzzaman pada tahun 2020 menyarankan sistem berbasis penglihatan yang mengenali karakter bahasa isyarat Arab dan menerjemahkannya ke dalam suara Arab menggunakan CNN. Keandalan ditunjukkan oleh akurasi pengenalan sistem sebesar 90% [5]. Dalam penelitian mereka tahun 2023, Izzalhaqqi dan Wahyono menciptakan model CNN-LSTM yang dapat mengenali gerakan alfabet statis

dalam Bahasa Isyarat Indonesia (BISINDO). Mereka menggunakan kumpulan data yang dikumpulkan sendiri dan dioptimalkan dengan Randomized Search CV untuk melakukan hal ini. Model CNN-LSTM menunjukkan hasil yang sangat baik, dengan skor rata-rata makro 0,98 di semua kelas dan akurasi pengujian 98,00%.

Namun, model VGG-16 dan CNN murni sedikit lebih baik daripada metode hibrida. [2] Untuk meningkatkan identifikasi Bahasa Isyarat Tiongkok, Han et al. melakukan studi kedua pada tahun 2025 yang menyajikan model STGCN-LSTM aliran ganda yang menggabungkan ciri-ciri fonologis seperti bentuk tangan, posisi, orientasi, dan gerakan dengan variabel rangka spasiotemporal. Dengan akurasi 95,2% pada dataset SRL500 dan akurasi 93,0% pada subset kata-kata yang terkait secara visual, metode mereka terbukti sangat akurat. [14]. Sementara penelitian ekstensif telah dilakukan pada ASL dan BISINDO, Bahasa Isyarat Argentina (LSA) masih kurang dieksplorasi. Hal ini memotivasi pengembangan alat otomatis yang dapat mendukung pengguna LSA melalui sistem pengenalan bahasa isyarat berdasarkan data video.

CNN sangat baik dalam mengekstraksi karakteristik spasial dari gambar, tetapi tidak terlalu baik dalam mengumpulkan korelasi temporal antara urutan bingkai. Jaringan LSTM, di sisi lain, sebagian besar digunakan untuk menangani data sekuensial seperti deret waktu, teks, atau audio. Namun, karena gambar berdimensi tinggi dan mengandung pola spasial yang rumit, jaringan LSTM tidak cocok untuk memprosesnya secara langsung. Video adalah jenis data visual sekuensial yang terdiri dari beberapa bingkai gambar yang menyampaikan perkembangan waktu dan informasi spasial. Seringkali diperlukan untuk menggabungkan arsitektur CNN dan LSTM untuk memproses data tersebut secara efisien. Metode hibrida ini, juga dikenal sebagai LSTM Konvolusional, menggabungkan keunggulan kedua model: Sementara lapisan LSTM merekam dinamika temporal di seluruh urutan bingkai, lapisan CNN digunakan untuk mengekstraksi karakteristik spasial dari setiap bingkai. [11] Pemodelan pola spasiotemporal yang lebih tepat dimungkinkan oleh jaringan LSTM konvolusional, yang, berbeda dengan model LSTM klasik, dapat menerima masukan gambar sambil mempertahankan struktur spasialnya. Aplikasi seperti pengenalan gestur, deteksi tindakan, dan kategorisasi video sangat diuntungkan dari integrasi ini karena memudahkan pemahaman data video secara keseluruhan. Untuk menciptakan model yang dapat mempelajari dimensi spasial dan temporal masukan video secara efisien, CNN dan LSTM harus digabungkan [11].

Penelitian ini berfokus pada pembangunan dan evaluasi model hibrida CNN-LSTM untuk mengenali 64 gestur Bahasa Isyarat Argentina menggunakan dataset LSA64. Berbeda dengan Cued Speech, yang menggabungkan isyarat tangan dengan membaca bibir, penelitian ini secara eksklusif membahas gestur bahasa isyarat yang mewakili kata atau ekspresi dalam LSA. Tujuan utamanya adalah untuk menunjukkan efektivitas model pembelajaran mendalam spasiotemporal hibrida dalam mengenali bahasa isyarat dari data video. Maksud dari kata di atas dalam bahasa Indonesia.