

BAB I

PENDAHULUAN

1.1 Latar Belakang

Komunikasi adalah aspek fundamental dalam kehidupan manusia, memungkinkan kita untuk berbagi informasi, mengekspresikan emosi, dan membangun hubungan sosial. Bagi penyandang tunarungu, komunikasi berjalan dengan menggunakan bahasa isyarat, yaitu bahasa yang memanfaatkan isyarat visual gerakan tangan dan tubuh, juga ekspresi wajah, untuk menyampaikan sesuatu selayaknya berkomunikasi (Alnabih & Maghari, 2024). Bahasa isyarat memberikan jembatan komunikasi antara penyandang tunarungu dan orang-orang di sekitar mereka, terutama dengan anggota keluarga, teman, rekan kerja, dan membantu mereka dalam memperoleh pendidikan. Banyak sekolah khusus untuk membantu anak-anak tunarungu belajar, berinteraksi dengan guru, dan mengembangkan keterampilan kognitif mereka, dengan menggunakan bahasa isyarat. Setiap gerakan dalam bahasa isyarat merepresentasikan sebuah huruf, kata, atau angka, yang kemudian dirangkai menjadi sebuah kalimat (Attia dkk., 2023). Uniknya, bahasa isyarat ternyata memiliki sejarah yang kaya dan telah berkembang menjadi cara berkomunikasi yang berbeda-beda antar wilayah dan negara, di mana masing-masing memiliki tata bahasa dan kosakatanya sendiri (Y. Zhang & Jiang, 2024).

Menurut Federasi Tunarungu Dunia (*World Federation of the Deaf*), ada lebih dari 300 bahasa isyarat yang diakui di seluruh dunia dan telah digunakan oleh 70 juta orang tuli (Rastgoo dkk., 2021). Dari banyaknya jenis bahasa isyarat di seluruh dunia, yang dipakai secara umum di Indonesia adalah Bahasa Isyarat Indonesia, atau dikenal sebagai BISINDO (Indra dkk., 2019). BISINDO telah menjadi pintu menuju kehidupan sosial, pendidikan, dan ekonomi yang lebih baik, bagi penyandang disabilitas tunarungu (Sari, 2023). Upaya untuk meningkatkan aksesibilitas bagi mereka dalam mengakses kebutuhan sosial juga terus berkembang. Kebijakan-kebijakan yang mendukung hak penyandang disabilitas agar lebih bisa bersosialisasi terus dibangun (Dwijayanti dkk., 2021). Dalam

upaya mendukung peningkatan aksesibilitas ini, pemanfaatan teknologi dalam pendeteksian bahasa isyarat serta penerjemah gerakan bahasa isyarat menjadi sangat diperlukan (Hao dkk., 2024). Namun hingga kini ketersediaan dataset yang representatif untuk BISINDO masih sangat terbatas. Ketidaksediaan dataset yang variatif dan berkualitas menghambat pengembangan teknologi berbasis kecerdasan buatan, khususnya dalam pengenalan bahasa isyarat untuk mendukung komunikasi inklusif. Oleh karena itu, upaya untuk menciptakan dataset BISINDO yang terstruktur, variatif, dan mencakup berbagai pola isyarat sangatlah mendesak guna mendukung penelitian dan pengembangan di bidang ini.

Pengenalan bahasa isyarat visual adalah area penelitian yang kompleks dalam *computer vision*, karena variasi bentuk tangan, profil gerakan, dan posisi bagian tubuh (Mienye dkk., 2024) tetapi teknologi pengenalan bahasa isyarat telah mengalami perkembangan pesat dengan adanya metode *deep learning*. Dan sudah banyak model yang telah diusulkan oleh berbagai peneliti dengan peningkatan signifikan oleh pendekatan *deep learning* dalam beberapa tahun terakhir. Tren keseluruhan dari model yang diusulkan menunjukkan peningkatan signifikan dalam akurasi pengenalan bahasa isyarat (De Coster dkk., 2021). Hal ini disebabkan oleh *deep learning* yang memungkinkan pengenalan bahasa isyarat secara otomatis dari video. Teknologi ini dapat membantu dalam berbagai aplikasi, seperti pendidikan, layanan publik, dan komunikasi sehari-hari.

Model *deep learning*, telah menunjukkan hasil yang memuaskan dalam banyak dimensi implementasi (Ismail dkk., 2021). Banyak penelitian yang telah membuktikan bahwa penggunaan *deep learning* memberikan kinerja yang lebih baik daripada teknik *machine learning* konvensional lainnya, terutama di bidang *computer vision* dan *natural language processing* (Y. Zhang & Jiang, 2024). Selama beberapa tahun terakhir, *deep learning* telah memiliki pengaruh besar pada aktivitas yang berkaitan dengan video. Misalnya, *recurrent neural networks* (RNN) telah menunjukkan kinerja yang signifikan dalam mempelajari hubungan *temporal* dalam pengenalan tanda dan identifikasi tindakan manusia, serta pengenalan gerakan dan pengenalan tanda (Abdulhamied dkk., 2023).

Pengembangan RNN juga menandai kemajuan signifikan dalam *deep learning*. RNN menggabungkan perhitungan terbaru ke dalam perhitungan saat ini, yang secara signifikan meningkatkan hasil di sejumlah bidang, termasuk *language processing* (Khartheesvar dkk., 2024). Berbeda dengan jaringan saraf tiruan biasa, RNN memiliki kemampuan tambahan untuk menggunakan *input* sebelumnya dalam memprediksi *output*. Artinya, RNN dapat menyimpan dan memodelkan urutan *temporal* dari *input* yang sudah diberikan (Abdulhamied dkk., 2023). Namun, RNN memiliki kelemahan yaitu hanya menyimpan data sebelumnya dalam memori jangka pendek mereka. Informasi yang disimpan paling lama akan dihapus ketika memori penuh dan digantikan dengan informasi baru. Oleh karena itu, dikembangkanlah model *Long short-term memory* (LSTM) yang dapat mengatasi masalah ini dengan menyimpan beberapa data dalam *cell state* dan menyimpannya dalam memori jangka panjang. Data jangka pendek dari tahap perhitungan sebelumnya juga disimpan dalam *hidden state*.

Arsitektur LSTM, yang merupakan salah satu jenis RNN, mendukung proses *deep learning* dan *machine learning* dengan lebih baik. Tidak seperti jaringan saraf tiruan konvensional, LSTM memiliki koneksi umpan balik yang memungkinkan pemrosesan tidak hanya pada titik data individu (seperti gambar), tetapi juga pada seluruh urutan data (Tran dkk., 2023). LSTM dirancang untuk menangani masalah ini dengan menggunakan gerbang memori yang dapat menjaga informasi selama jangka waktu yang panjang. LSTM memiliki tiga gerbang utama yaitu *forget gate*, *input gate*, dan *output gate* yang memungkinkan model untuk memilih informasi mana yang harus dipertahankan atau dilupakan, sehingga dapat memproses ketergantungan jangka panjang secara efektif (Chen dkk., 2020).

Dalam pengembangan teknologi pengenalan bahasa isyarat, sangat diperlukan pemodelan hubungan spasial dan *temporal* berdasarkan data dari video, di mana LSTM, memiliki keunggulan karena mampu menyimpan informasi dari *frame* sebelumnya dan menggunakannya untuk memperbaiki prediksi pada *frame* saat ini (Li dkk., 2020). Model LSTM telah digunakan secara luas untuk menangani tugas-tugas yang memerlukan pemahaman urutan data, dan

menjadi salah satu alasan mengapa LSTM sangat efektif dalam pengenalan bahasa isyarat karena kemampuannya untuk menangkap ketergantungan *temporal* jangka panjang dari urutan video (Huang dkk., 2024), juga penambahan mekanisme *forget gate* dan *input-output gate* dalam LSTM membuatnya lebih stabil dalam menangani data yang panjang dan kompleks, seperti video gerakan tangan dalam bahasa isyarat (Li dkk., 2020).

Deep learning merupakan pendekatan yang dapat digunakan untuk melakukan pengenalan gambar, pemrosesan bahasa alami, deteksi objek, dan pengenalan bahasa isyarat, seperti RNN arsitektur LSTM yang digunakan untuk data sekuensial seperti video. Dalam konteks penelitian pengenalan BISINDO, model LSTM akan sangat cocok karena kemampuan LSTM dalam mempelajari urutan data yang penting, seperti gerakan tangan dalam video. Pendekatan ini memungkinkan pembangunan model yang secara langsung dapat menerjemahkan bahasa isyarat kedalam sebuah teks sesuai dengan gerakan tangannya, sehingga harapannya, teknologi ini dapat berpotensi untuk memecahkan hambatan komunikasi dan memungkinkan penyandang tunarungu untuk berinteraksi lebih luas di masyarakat.

Meskipun demikian, performa LSTM yang baik dalam menangani data sekuensial sering kali bergantung pada kualitas dan keragaman data yang digunakan, serta pada konfigurasi arsitekturnya. Dalam konteks pengenalan BISINDO, model LSTM perlu diperkuat (*enhanced*) agar dapat beradaptasi dengan kompleksitas gerakan tangan, ekspresi wajah, dan variasi spasial-temporal yang tinggi dalam data video. Tantangan seperti *vanishing gradient*, ketidakseimbangan data, dan kecenderungan *overfitting* pada jumlah data yang terbatas menjadi faktor yang menurunkan akurasi jika model tidak dioptimalkan. Oleh karena itu, diperlukan pendekatan penguatan model melalui teknik seperti penyesuaian *hyperparameter*, regularisasi, optimalisasi jumlah *neuron*, serta penyempurnaan proses pelatihan agar LSTM mampu belajar secara lebih stabil dan efisien. Upaya penguatan ini penting untuk memastikan bahwa model tidak hanya mampu mengenali pola gerakan secara umum, tetapi juga memahami nuansa dan variasi gerakan khas BISINDO secara lebih akurat.

Berdasarkan latar belakang tersebut, berkomunikasi dengan bahasa isyarat yang adalah hal yang umum digunakan. Namun, komunikasi menjadi sulit ketika salah satu pihak tidak menguasainya (Saqib & Rahman, 2020). Apalagi, jika dibandingkan dengan bahasa lisan, bahasa isyarat jauh lebih sulit dipelajari karena kompleksitas dan spesialisasinya yang tinggi, seperti gerakan tangan khusus, postur, dan ekspresi wajah (Enikeev & Mustafina, 2021). Dalam permasalahan ini, teknologi penerjemah untuk bahasa isyarat perlu dibangun, dan dengan munculnya teknologi yang dapat mengenali bahasa isyarat dapat membantu meningkatkan aksesibilitas mereka dalam melakukan komunikasi inklusif ke sosial masyarakat (Shinde dkk., 2024).

Pengumpulan dataset BISINDO yang berkualitas juga menjadi perhatian utama dalam membangun model pengenalan bahasa isyarat menggunakan *deep learning*. Dataset yang dibangun tidak hanya diperlukan untuk melatih model agar mampu mengenali berbagai isyarat secara akurat, tetapi juga berperan dalam mempromosikan inklusivitas dengan menyediakan solusi teknologi yang memungkinkan komunikasi dua arah antara komunitas tunarungu dan masyarakat umum. Dengan adanya dataset BISINDO yang terstandarisasi, penelitian lebih lanjut di bidang pengenalan bahasa isyarat dapat dilakukan secara lebih komprehensif dan efisien, membuka peluang implementasi teknologi yang lebih luas di berbagai sektor seperti pendidikan, layanan publik, dan interaksi sosial.

Disisilain, seiring berkembangnya teknologi *deep learning* dan *computer vision*, pengembangan model yang mampu mengenali dan menerjemahkan bahasa isyarat dari video secara langsung menjadi lebih memungkinkan. *long short-term memory* (LSTM) sebagai salah satu arsitektur dalam *deep learning recurrent neural networks* (RNN) telah terbukti efektif dalam menangani data sekuensial seperti video. Oleh karena itu, diperlukan penelitian lebih lanjut tentang bagaimana model *deep learning*, khususnya LSTM, dapat digunakan untuk mengenali BISINDO dari video, ditambah lagi dari keseluruhan penelitian terkait, sebagian besar penelitian pengenalan bahasa isyarat saat ini masih berfokus pada bahasa isyarat internasional, seperti *American Sign language* (ASL), sementara penelitian yang berfokus pada BISINDO masih sangat terbatas.

Oleh karena itu, mengimplementasikan teknologi *deep learning* berbasis *recurrent neural network* (RNN) dengan arsitektur LSTM dalam membangun model penerjemah BISINDO menjadi suatu tantangan baru. Model yang akan diimplementasikan diharapkan dapat menjadi nilai tambah dalam pengembangan teknologi ini, sehingga menghasilkan intervensi yang lebih akurat dan efektif dalam upaya mendukung pengembangan teknologi penerjemah bahasa isyarat khususnya di Indonesia. Penelitian ini juga diharapkan dapat mengembangkan dataset BISINDO terbaru, yang variatif dan memungkinkan penelitian lanjutan di masa mendatang, yang dapat mengatasi kesenjangan komunikasi antara pengguna BISINDO dan masyarakat umum.

1.2 Tujuan Penelitian

Tujuan penelitian ini adalah mengembangkan model *deep learning* berbasis *Long short-term memory* (LSTM) untuk mengenali dan menerjemahkan gerakan tangan, wajah, dan pose tubuh dalam Bahasa Isyarat Indonesia (BISINDO) secara akurat. Selain itu, penelitian ini bertujuan menyusun dan memperkaya dataset BISINDO yang merepresentasikan variasi gerakan isyarat secara memadai sebagai dasar pelatihan model. Penelitian ini juga mengevaluasi kinerja model pada berbagai variasi gerakan untuk memastikan keandalan dan konsistensi performa dalam kondisi penggunaan yang beragam.

1.3 Manfaat Penelitian

Berdasarkan tujuan yang ada, penelitian ini diharapkan memberikan berbagai manfaat, terutama dalam memperkuat peran teknologi informasi dalam mendukung aksesibilitas dan inklusi sosial bagi publik. Hasil penelitian ini akan memberikan kontribusi penting bagi perkembangan teknologi pengenalan bahasa isyarat otomatis, khususnya untuk BISINDO, yang sebelumnya kurang mendapat perhatian dibandingkan dengan bahasa isyarat internasional lainnya seperti ASL. Model yang dikembangkan dalam penelitian ini juga berpotensi berfungsi sebagai alat bantu komunikasi yang mampu menerjemahkan isyarat BISINDO ke dalam teks secara langsung, sehingga mempermudah interaksi antara pengguna dan non-pengguna BISINDO dalam berbagai situasi sehari-hari. Selain itu, penelitian ini dapat menghasilkan *prototipe* aplikasi berbasis *deep learning* dan *computer vision*

yang mampu mengenali serta menerjemahkan gerakan BISINDO secara akurat. Aplikasi ini nantinya dapat diimplementasikan pada berbagai perangkat seperti *smartphone* atau tablet untuk mendukung komunikasi komunitas tunarungu. Dengan memperkaya dataset gerakan BISINDO melalui pengumpulan data baru, penelitian ini juga membuka peluang bagi studi lanjutan serta menyediakan dataset yang lebih lengkap dan bervariasi guna mendukung penelitian di masa depan.

1.4 Batasan penelitian

Penelitian ini memiliki beberapa batasan yang ditetapkan agar ruang lingkup penelitian tetap terarah dan sesuai dengan tujuan yang telah dirumuskan. Penelitian ini berfokus secara khusus pada BISINDO dan tidak mencakup bahasa isyarat lain seperti *American Sign Language* (ASL) maupun Sistem Isyarat Bahasa Indonesia (SIBI). Data yang digunakan berupa video gerakan isyarat yang telah diproses menjadi data sekuens berbasis landmark melalui ekstraksi menggunakan *MediaPipe Holistic*, dengan urutan fitur landmark terbatas di wajah, tangan, lalu pose. Penelitian ini tidak memanfaatkan data citra mentah maupun sensor tambahan seperti depth sensor. Dataset yang digunakan terbatas pada sejumlah kelas tertentu, yaitu sekitar 76 kelas gerakan, sehingga belum merepresentasikan keseluruhan kosakata dalam BISINDO. Selain itu, setiap data diproses dalam bentuk sekuens dengan panjang tetap sebesar 30 *frame* per video, sehingga variasi durasi gerakan di luar skenario tersebut tidak menjadi fokus analisis. Fitur yang digunakan dalam penelitian ini juga dibatasi pada *landmark* pose, tangan, dan wajah dengan total 1662 fitur per *frame*, tanpa mempertimbangkan fitur lain seperti *optical flow* atau informasi berbasis piksel mentah. Dari sisi pemodelan, penelitian ini hanya menggunakan arsitektur LSTM yang telah ditingkatkan (*enhanced*).