

Pengembangan Dataset Video Gerakan Bibir untuk Pengenalan Kata Bahasa Indonesia

Friska Adisti Mahardini*¹, Dr. Ir. Anang Kukuh Adisusilo, ST., MT.²

^{1,2} Universitas Wijaya Kusuma Surabaya, Surabaya, Jawa Timur, Indonesia

Email: ¹friskamahardini@gmail.com, ²anang65@uwks.ac.id

*Penulis Korespondensi

Abstrak

Dataset video gerakan bibir merupakan komponen penting dalam penelitian *visual speech recognition* karena pengenalan kata berbasis visual membutuhkan data yang mampu merepresentasikan pola artikulasi secara jelas. Kebutuhan tersebut menjadi tantangan pada bahasa Indonesia karena dataset video gerakan bibir masih perlu diperluas, terutama untuk pengenalan kata dengan data yang dikumpulkan secara terkontrol. Penelitian ini bertujuan untuk mengembangkan dataset video gerakan bibir untuk pengenalan kata bahasa Indonesia yang diharapkan dapat menjadi dasar awal bagi penelitian lanjutan dalam pengembangan sistem pengenalan kata berbasis informasi visual. Metode yang digunakan meliputi penentuan kelas kata, perekaman video, pengorganisasian data berdasarkan label, pra-pemrosesan frame, dan validasi awal menggunakan model berbasis *deep learning*. Dataset yang dihasilkan terdiri dari 796 video dengan 12 kelas kata, dengan setiap video memuat satu kata yang direkam pada posisi wajah frontal. Hasil validasi awal menunjukkan akurasi uji sebesar 96,25%, sehingga dataset yang dikembangkan dapat digunakan sebagai dasar penelitian pengenalan kata bahasa Indonesia berbasis gerakan bibir.

Kata kunci: Bahasa Indonesia, dataset, gerakan bibir, pengenalan kata, *visual speech recognition*.

Abstract

Lip movement video datasets are an important component in visual speech recognition research because visual word recognition requires data that can clearly represent articulation patterns. This need becomes a challenge for Indonesian because lip movement video datasets still need to be expanded, especially for word recognition using data collected in a controlled manner. This study aims to develop a lip movement video dataset for Indonesian word recognition. The dataset is expected to serve as an initial foundation for further research in developing word recognition systems based on visual information. The method includes word class selection, video recording, label-based data organization, frame preprocessing, and initial validation using a deep learning-based model. The resulting dataset consists of 796 videos with 12 words classes, where each video contains one word recorded in a frontal face position. The initial validation result shows a test accuracy of 96.25%, indicating that the developed dataset can be used as a foundation for Indonesian word recognition research based on lip movements.

Keywords: dataset, Indonesian language, lip movement, visual speech recognition, word recognition.

I. PENDAHULUAN

Visual speech recognition (VSR) merupakan bidang penelitian yang berfokus pada pengenalan ucapan berdasarkan informasi visual dari gerakan bibir. Pendekatan ini tidak bergantung sepenuhnya pada sinyal audio, sehingga dapat digunakan pada kondisi ketika suara tidak tersedia, terganggu, atau memiliki kualitas rendah. Dalam konteks pengenalan kata, gerakan bibir menjadi sumber informasi visual karena setiap kata memiliki pola artikulasi yang dapat diamati melalui perubahan bentuk mulut. Penelitian *VSR* juga semakin berkembang seiring dengan kemajuan metode *deep learning* dan ketersediaan dataset audio-visual yang mendukung proses pelatihan model [1] [2].

Dataset menjadi komponen penting dalam penelitian *VSR* karena model pembelajaran mendalam membutuhkan data visual yang mampu merepresentasikan variasi gerakan bibir. Ketersediaan dataset yang sesuai dengan bahasa target berpengaruh terhadap pengembangan sistem pengenalan kata berbasis visual. Hal ini terjadi karena pola artikulasi, kosakata, dan karakteristik pengucapan dapat berbeda pada setiap bahasa. Oleh karena itu, pengembangan dataset pada bahasa tertentu diperlukan agar penelitian *VSR* tidak hanya bergantung pada dataset dari bahasa lain.

Beberapa dataset *VSR* telah dikembangkan untuk mendukung penelitian lip reading pada berbagai bahasa. GLips merupakan dataset bahasa Jerman yang terdiri dari 250.000 video wajah pembicara dan diproses untuk *word-level lip reading*. Dataset tersebut dikembangkan karena dataset besar untuk pembelajaran mendalam pada lip reading belum tersedia di banyak bahasa [3]. Pada konteks bahasa

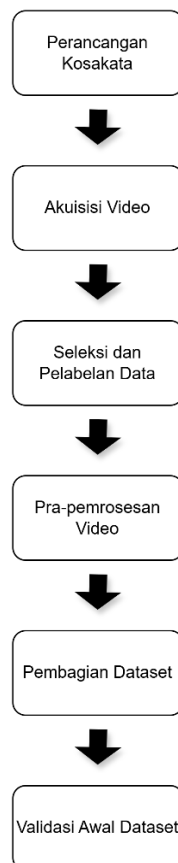
Indonesia, *IDLRW* menyediakan 48.083 video dari 100 kata bahasa Indonesia, sedangkan *ID-LR* menyediakan 105.300 video dari 230 kata bahasa Indonesia. Kedua dataset tersebut menunjukkan bahwa penelitian lip reading bahasa Indonesia mulai berkembang dan membutuhkan dukungan data visual yang semakin beragam [4][5].

Meskipun dataset *lip reading* bahasa Indonesia telah mulai tersedia, pengembangan dataset berskala awal dengan proses perekaman terkontrol tetap relevan untuk mendukung eksperimen dasar. Dataset terkontrol dapat digunakan untuk mengamati pola gerakan bibir pada kata tertentu, menyusun data berdasarkan label secara jelas, dan menyediakan data awal untuk validasi metode pengenalan kata. Pendekatan ini penting terutama pada penelitian yang diarahkan sebagai tahap awal sebelum pengembangan sistem pengenalan gerakan bibir yang lebih kompleks.

Penelitian ini bertujuan untuk mengembangkan dataset video gerakan bibir untuk pengenalan kata bahasa Indonesia. Dataset yang dikembangkan terdiri dari 796 video dengan 12 kelas kata, yaitu apel, bantu, bola, halo, haus, jeruk, lapar, makan, minum, saya, tolong, dan tomat. Setiap video memuat satu kata yang direkam dengan posisi wajah frontal. Dataset dikembangkan melalui penentuan kelas kata, perekaman video, pengorganisasian data berdasarkan label, pra-pemrosesan frame, dan validasi awal menggunakan model *CNN-LSTM*. Kontribusi utama penelitian ini adalah menyediakan dataset video gerakan bibir yang dapat digunakan sebagai dasar awal untuk penelitian pengenalan kata bahasa Indonesia berbasis informasi visual.

II. METODE PENELITIAN

Metode penelitian ini berfokus pada proses pengembangan dataset video gerakan bibir untuk pengenalan kata bahasa Indonesia. Tahapan penelitian dilakukan secara berurutan mulai dari perancangan kosakata, akuisisi video gerakan bibir, seleksi dan pelabelan data, pra-pemrosesan video, pembagian dataset, hingga validasi awal dataset. Setiap tahap dirancang untuk menghasilkan dataset yang terstruktur, memiliki label yang jelas, dan dapat digunakan sebagai dasar penelitian pengenalan kata berbasis informasi visual. Alur tahapan penelitian ditunjukkan pada Gambar 1.



Gambar 1. Alur Metode Penelitian

2.1. Perancangan Kosakata

Dataset pada penelitian ini dirancang pada tingkat kata untuk mendukung pengenalan kata bahasa Indonesia berbasis gerakan bibir. Setiap data video hanya memuat satu kata sehingga pola gerakan bibir pada setiap label dapat diamati secara lebih terkontrol. Pendekatan tingkat kata dipilih sebagai tahap awal karena proses perekaman, pelabelan, dan validasi dataset dapat dilakukan dengan lebih jelas sebelum dikembangkan ke bentuk pengenalan frasa atau kalimat.

2.2. Akusisi Video

Akuisisi video dilakukan dengan merekam gerakan bibir saat pengucapan kata bahasa Indonesia yang telah ditentukan pada tahap perancangan kosakata. Setiap video hanya memuat satu kata sehingga data yang diperoleh memiliki hubungan langsung dengan label kelas yang digunakan. Proses perekaman dilakukan secara mandiri menggunakan kamera laptop. Video direkam dengan resolusi 720p dan frame rate 30 fps. Durasi setiap video berada pada rentang sekitar 1–2 detik. Selama perekaman, posisi wajah diarahkan secara frontal agar area bibir terlihat jelas. Perekaman dilakukan pada kondisi pencahayaan dalam ruangan dan luar ruangan untuk memperoleh variasi visual pada data.

2.3. Seleksi dan Pelabelan Data

Seleksi data dilakukan untuk memastikan bahwa video yang digunakan memiliki kualitas visual yang cukup untuk merepresentasikan gerakan bibir. Video yang dipertahankan adalah video yang memuat satu kata, memperlihatkan area bibir dengan jelas, dan sesuai dengan label kata yang telah ditentukan. Data yang tidak memenuhi kriteria tersebut tidak digunakan dalam dataset akhir. Setiap video diberi label berdasarkan kata yang diucapkan. Pelabelan dilakukan dengan mengelompokkan video ke dalam folder sesuai kelas kata, sehingga setiap folder merepresentasikan satu label. Struktur ini digunakan untuk memudahkan proses pembacaan data, pemeriksaan jumlah data setiap kelas, dan tahap pra-pemrosesan.

2.4. Pra-pemrosesan Video

Pra-pemrosesan video dilakukan untuk menghasilkan data masukan yang seragam dari setiap video gerakan bibir. Setiap video dibaca menggunakan *OpenCV*, kemudian *frame* video dikonversi dari format warna *BGR* ke *RGB*. Konversi ini dilakukan agar format warna frame sesuai dengan format visual yang digunakan. Setelah video dibaca, setiap video diambil sebanyak 25 *frame* menggunakan teknik pengambilan *frame* secara seragam atau *uniform sampling*. Teknik ini mengambil *frame* dari posisi yang tersebar merata sepanjang durasi video, sehingga video dengan jumlah *frame* berbeda tetap menghasilkan jumlah *frame* yang sama. Jika sebuah video memiliki total *frame* sebanyak N , dan jumlah *frame* yang akan diambil adalah $K=25$, maka indeks *frame* ke- i dapat dihitung menggunakan Persamaan 1.

$$f_i = \left\lceil \frac{i * (N - 1)}{K - 1} \right\rceil, i = 0, 1, 2, 3, \dots, K - 1 \quad [1]$$

Tahap berikutnya adalah peningkatan kontras menggunakan metode *Contrast Limited Adaptive Histogram Equalization (CLAHE)*. *CLAHE* diterapkan pada kanal L dalam ruang warna LAB untuk meningkatkan kontras pencahayaan pada area bibir tanpa mengubah informasi warna secara langsung. Setelah itu, nilai piksel dinormalisasi ke rentang 0–1. Hasil akhir dari tahap ini adalah sekumpulan *frame* terstandar yang dapat digunakan pada proses validasi awal dataset.

2.5. Pembagian Dataset

Dataset yang telah melalui tahap pra-pemrosesan dibagi menjadi data latih dan data uji. Pembagian dilakukan menggunakan rasio 80:20, sehingga sebagian besar data digunakan untuk proses pelatihan pada tahap validasi awal, sedangkan sisanya digunakan untuk menguji kemampuan model dalam mengenali data yang tidak digunakan saat pelatihan. Pembagian dataset dilakukan menggunakan fungsi *train_test_split* dengan parameter *test_size*=0.2 dan *random_state*=42. Berdasarkan pembagian tersebut, diperoleh 636 video sebagai data latih dan 160 video sebagai data uji.

2.6. Validasi Awal Dataset

Validasi awal dilakukan untuk memastikan bahwa dataset yang dikembangkan dapat digunakan pada tugas pengenalan kata berbasis gerakan bibir. Pada tahap ini, model berbasis *deep learning* digunakan sebagai alat uji untuk melihat keterpakaian dataset dalam proses klasifikasi. Validasi ini tidak diarahkan

untuk membandingkan arsitektur model, tetapi untuk menunjukkan bahwa data yang dikembangkan memiliki pola visual yang dapat dipelajari oleh model.

III. HASIL DAN PEMBAHASAN

3.1. Hasil Perancangan Kosakata

Hasil dari tahap perancangan kosakata adalah daftar kelas kata yang digunakan sebagai dasar pengembangan dataset. Kosakata yang digunakan terdiri dari 12 kata bahasa Indonesia yang umum digunakan dalam komunikasi sehari-hari. Setiap kata diperlakukan sebagai satu kelas dalam dataset, sehingga setiap video memiliki label yang jelas sesuai kata yang diucapkan. Daftar kelas kata yang digunakan ditunjukkan pada Tabel 1.

Tabel 1. Daftar Kelas Kata Dataset

No.	Kelas Kata
1.	Apel
2.	Bantu
3.	Bola
4.	Halo
5.	Haus
6.	Jeruk
7.	Lapar
8.	Makan
9.	Minum
10.	Saya
11.	Tolong
12.	Tomat

3.2. Hasil Akuisisi dan Seleksi Data

Tahap akuisisi data menghasilkan 960 video gerakan bibir dari seluruh kelas kata yang telah ditentukan. Data tersebut diperoleh melalui proses perekaman secara mandiri. Setelah seluruh data video terkumpul, proses seleksi dilakukan untuk memastikan data yang digunakan memiliki kesesuaian antara isi video, label kelas, dan kualitas visual area bibir. Data video yang dipertahankan adalah data yang memuat satu kata sesuai kosakata, memiliki tampilan area bibir yang jelas, dan tidak mengalami gangguan visual yang dapat menghambat tahap pra-pemrosesan. Data yang tidak memenuhi kriteria tersebut tidak digunakan dalam dataset akhir.

Setelah proses seleksi, jumlah data akhir yang diperoleh adalah 796 video. Sebanyak 164 video lainnya tidak digunakan karena tidak memenuhi kriteria seleksi. Data yang telah terpilih kemudian dikelompokkan berdasarkan label kelas agar dapat diolah secara lebih terstruktur. Ringkasan hasil akuisisi dan seleksi dataset ditunjukkan pada Tabel 2.

Tabel 2. Hasil Akuisisi dan Seleksi Dataset

Komponen	Jumlah	Keterangan
Data awal hasil akuisisi	960 video	Seluruh video yang terkumpul dari proses perekaman
Data akhir setelah seleksi	796 video	Video yang memenuhi kriteria visual dan label
Data tidak digunakan	164 video	Video yang tidak memenuhi kriteria seleksi

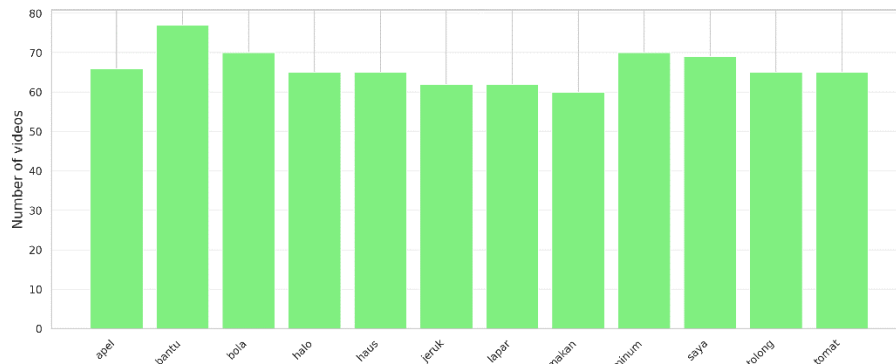
3.3. Distribusi Dataset

Distribusi dataset menunjukkan jumlah video pada setiap kelas kata setelah proses seleksi data dilakukan. Informasi ini digunakan untuk melihat komposisi data pada masing-masing kelas dan memastikan bahwa setiap kelas memiliki data yang dapat digunakan pada tahap pra-pemrosesan serta validasi awal. Distribusi jumlah video pada setiap kelas ditunjukkan pada Tabel 3 dan divisualisasikan pada Gambar 2.

Tabel 3. Hasil Akuisisi dan Seleksi Dataset

No.	Kelas Kata	Jumlah Video
1.	Apel	66 video
2.	Bantu	77 video
3.	Bola	70 video
4.	Halo	65 video
5.	Haus	65 video

6.	Jeruk	62 video
7.	Lapar	62 video
8.	Makan	60 video
No.	Kelas Kata	Jumlah Video
9.	Minum	70 video
10.	Saya	69 video
11.	Tolong	65 video
12.	Tomat	65 video



Gambar 2. Distribusi Data Setiap Kelas

Berdasarkan distribusi data, setiap kelas memiliki jumlah video yang berbeda. Perbedaan jumlah data antar-kelas terjadi karena proses seleksi dilakukan berdasarkan kualitas visual, kesesuaian label, dan kejelasan area bibir pada setiap video. Meskipun terdapat variasi jumlah data pada setiap kelas, seluruh kelas tetap memiliki data yang digunakan dalam proses validasi awal dataset.

3.4. Hasil Pra-pemrosesan Video

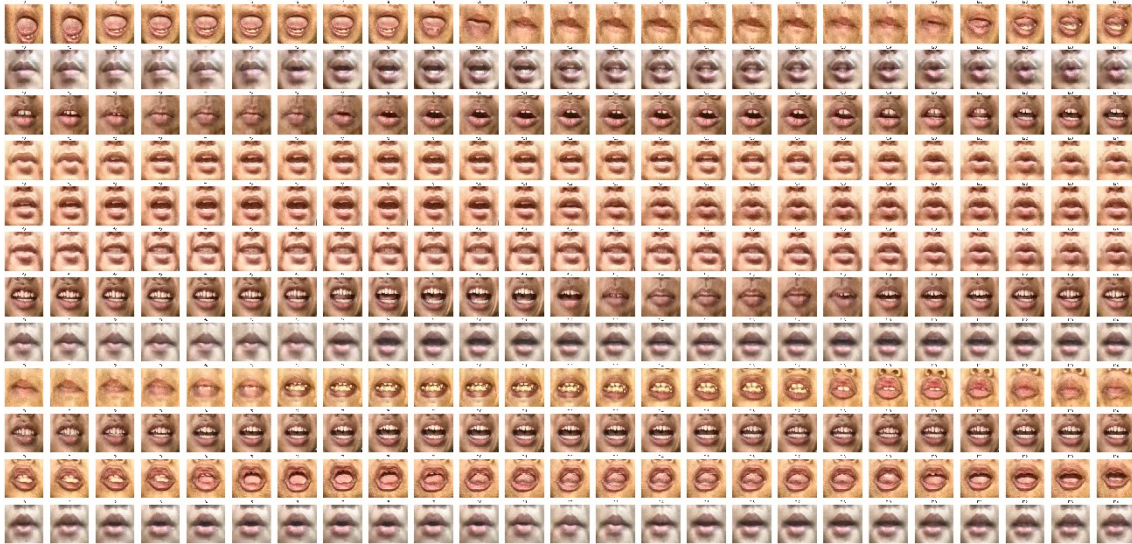
Pra-pemrosesan video menghasilkan data *frame* yang lebih seragam untuk seluruh video dalam dataset. Setiap video direpresentasikan menjadi 25 *frame* melalui teknik *uniform sampling*. Hasil ini membuat setiap video memiliki jumlah *frame* yang sama meskipun durasi video asli berada pada rentang yang berbeda. Setelah proses pengambilan *frame*, setiap *frame* diubah ukurannya menjadi 112×112 piksel. Perubahan ukuran ini menghasilkan dimensi spasial yang konsisten pada seluruh data.

Selain itu, peningkatan kontras menggunakan metode *Contrast Limited Adaptive Histogram Equalization (CLAHE)* menghasilkan tampilan area bibir yang lebih jelas, terutama pada bagian yang memiliki perbedaan pencahayaan. Nilai piksel kemudian dinormalisasi ke rentang 0–1 agar seluruh *frame* memiliki skala nilai yang seragam.

Tabel 4. Ringkasan Hasil Pra-pemrosesan Video

Komponen	Hasil
Jumlah frame per video	25 <i>frame</i>
Ukuran frame	112×112 piksel
Format warna	RGB
Peningkatan kontras	CLAHE
Normalisasi	Rentang 0–1

Tabel 4 menunjukkan bahwa proses pra-pemrosesan menghasilkan data video dengan jumlah frame, ukuran citra, format warna, dan rentang nilai piksel yang seragam. Keceragaman ini diperlukan agar data dapat digunakan secara konsisten pada tahap validasi awal dataset.



Gambar 3. Hasil Frame Sesudah Dilakukan Pra-pemrosesan

Contoh hasil frame setelah melalui tahap pra-pemrosesan ditunjukkan pada Gambar 3. Frame hasil pra-pemrosesan memiliki ukuran yang seragam dan kontras area bibir yang lebih jelas, sehingga pola gerakan bibir dapat diamati dengan lebih baik.

3.5. Hasil Pembagian Dataset

Dataset yang telah melalui tahap pra-pemrosesan dibagi menjadi data latih dan data uji menggunakan rasio 80:20. Hasil pembagian menghasilkan 636 video sebagai data latih dan 160 video sebagai data uji. Pembagian ini digunakan untuk mendukung tahap validasi awal dataset, sehingga sebagian data digunakan untuk proses pelatihan dan sebagian lainnya digunakan untuk pengujian.

Tabel 5. Hasil Pembagian Dataset

Jenis Data	Jumlah Video	Persentase
Data latih	636 Video	80%
Data Uji	160 Video	20%
Total	796 Video	100%

Tabel 5 menunjukkan bahwa sebagian besar data digunakan sebagai data latih, sedangkan data uji digunakan untuk mengevaluasi kemampuan model dalam mengenali data yang tidak digunakan pada proses pelatihan. Pembagian ini menjaga alur validasi awal dataset agar proses pelatihan dan pengujian dilakukan pada data yang berbeda.

3.6. Hasil Validasi Awal Dataset

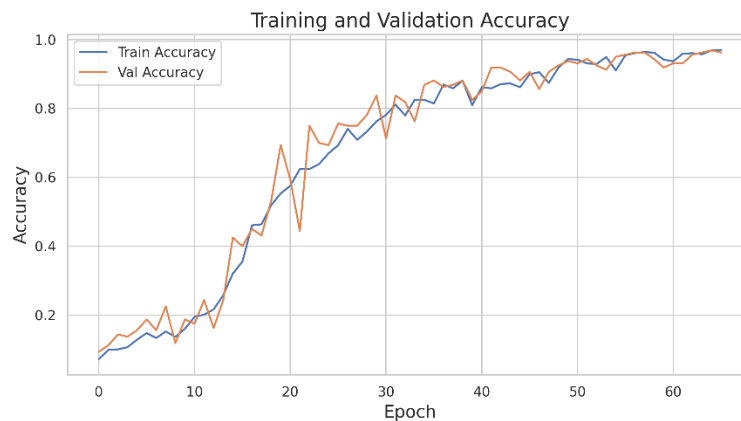
Validasi awal dilakukan untuk mengetahui apakah dataset yang dikembangkan dapat digunakan pada tugas pengenalan kata berbasis gerakan bibir. Pada tahap ini, model berbasis *deep learning* digunakan sebagai alat uji untuk mempelajari pola visual pada data hasil pra-pemrosesan. Validasi ini tidak diarahkan untuk membandingkan arsitektur model, tetapi untuk menilai kelayakan dataset dalam mendukung proses klasifikasi awal.

Hasil validasi awal menunjukkan bahwa dataset dapat digunakan dalam proses klasifikasi kata. Model memperoleh akurasi sebesar 97,01% pada data latih dan 96,25% pada data uji. Nilai loss pada data uji sebesar 0,144. Ringkasan hasil validasi awal ditunjukkan pada Tabel 6.

Tabel 6. Hasil Validasi Awal Dataset

Metrik	Nilai
Akurasi Data latih	97,01%
Akurasi Data Uji	96,25%
Loss Data Uji	0,144

Tabel 6 menunjukkan bahwa model mampu mempelajari pola visual dari dataset yang dikembangkan. Selisih akurasi antara data latih dan data uji tidak terlalu besar, sehingga hasil validasi awal menunjukkan bahwa dataset memiliki pola visual yang dapat digunakan untuk pengenalan kata berbasis gerakan bibir.



Gambar 4. Grafik Akurasi Pelatihan dan Validasi

Gambar 4 menunjukkan bahwa akurasi pelatihan dan validasi mengalami peningkatan selama proses pelatihan. Pola ini menunjukkan bahwa model dapat mempelajari representasi visual dari data yang telah diproses.



Gambar 4. Grafik Loss Pelatihan dan Validasi

Gambar 5 menunjukkan bahwa nilai loss cenderung menurun selama proses pelatihan. Penurunan loss menunjukkan bahwa kesalahan prediksi semakin berkurang pada proses validasi awal dataset.

Secara keseluruhan, hasil validasi awal menunjukkan bahwa dataset yang dikembangkan memiliki pola visual yang dapat dipelajari oleh model berbasis *deep learning*. Hasil ini mendukung bahwa dataset video gerakan bibir yang dihasilkan dapat digunakan sebagai dasar awal untuk penelitian pengenalan kata bahasa Indonesia. Meskipun demikian, validasi ini masih bersifat awal sehingga pengembangan lebih lanjut tetap diperlukan, terutama dengan jumlah kosakata, jumlah pembicara, dan variasi kondisi perekaman yang lebih luas.

IV. KESIMPULAN

Penelitian ini bertujuan untuk mengembangkan dataset video gerakan bibir untuk pengenalan kata bahasa Indonesia. Berdasarkan hasil pengembangan dataset, tujuan tersebut telah tercapai melalui penyusunan dataset yang terdiri dari 796 video dengan 12 kelas kata, yaitu Apel, Bantu, Bola, Halo, Haus, Jeruk, Lapar, Makan, Minum, Saya, Tolong, dan Tomat. Dataset dikembangkan melalui tahapan perancangan kosakata, akuisisi video gerakan bibir, seleksi dan pelabelan data, pra-pemrosesan video, pembagian dataset, serta validasi awal dataset.

Hasil pada bagian Hasil dan Pembahasan menunjukkan bahwa dataset yang dikembangkan memiliki struktur data yang jelas, label yang sesuai, serta data hasil pra-pemrosesan yang seragam. Setiap video direpresentasikan menjadi 25 *frame* berukuran 112 x 112 piksel dengan nilai piksel yang dinormalisasi ke rentang 0–1. Validasi awal menggunakan model berbasis *deep learning* menghasilkan akurasi uji sebesar 96,25%. Hasil tersebut menunjukkan bahwa dataset dapat digunakan sebagai dasar awal untuk penelitian pengenalan kata bahasa Indonesia berbasis gerakan bibir.

Pengembangan selanjutnya dapat dilakukan dengan memperluas jumlah kosakata, menambah jumlah pembicara, serta meningkatkan variasi kondisi perekaman, seperti pencahayaan, jarak kamera, dan latar belakang. Selain itu, dataset yang telah dikembangkan dapat digunakan sebagai dasar untuk penelitian lanjutan dalam pengembangan sistem pengenalan gerakan bibir pada tingkat frasa atau kalimat.

REFERENSI

- [1] P. Ma, S. Petridis, and M. Pantic, “Visual speech recognition for multiple languages in the wild,” *Nat. Mach. Intell.*, vol. 4, no. 11, pp. 930–939, 2022, doi: 10.1038/s42256-022-00550-z.
- [2] J. S. Chung and A. Zisserman, “Lip Reading in the Wild”.
- [3] G. Schwiebert, C. Weber, L. Qu, H. Siqueira, and S. Wermter, “A Multimodal German Dataset for Automatic Lip Reading Systems and Transfer Learning,” *2022 Lang. Resour. Eval. Conf. Lr. 2022*, no. June, pp. 6829–6836, 2022, doi: 10.25592/uhhfdm.10047.
- [4] G. Megiyanto Rahmatullah, “Indonesian Lipreading Dataset (IDLRW),” Dec. 2023, *Zenodo*. doi: 10.5281/zenodo.10432663.
- [5] G. Megiyanto Rahmatullah, “Indonesian Lip Reading (ID-LR),” Jul. 2024, *Zenodo*. doi: 10.5281/zenodo.12770253.