J-COSINE (Journal of Computer Science and Informatics Engineering)
Vol. 7, No. 2, December 2023 E-ISSN:2541-0806
Accredited Sinta-4 by RISTEKDIKTI Decree No. 79/E/KPT/2023 P-ISSN:2540-8895

Analisis Kebutuhan Dataset Algoritma Speech to Text Bahasa Sasak Menggunakan Perbandingan Data Suara Bahasa Inggris Pada Metode CNN

Analysis of Sasak Language Speech to Text Algorithm Dataset Requirements Using English Voice Data Comparison on CNN Method

Widya Bayu Pratiwi, Arik Aranta*, Gibran Satya Nugraha
Dept Informatics Engineering, Mataram University
Jl. Majapahit 62, Mataram, Lombok NTB, INDONESIA
Email: aydiwbayu@gmail.com, [arikaranta, gibransn]@unram.ac.id

Penulis Korespondensi

Abstract **Currently, there have been many studies on speech recognition or speech to text. Speech to text is a technology used to convert human speech or voice and translate it into written text. Some speech to text research that has been done, has obtained an accuracy rate of up to 95% with English datasets using the Mel Frequency Coefficient (MFCC) feature extraction method and the Convolutional Neural Network (CNN) classification method. This research will apply similar algorithms, namely MFCC and CNN by displaying the training process and the resulting accuracy in its processing with an analysis scenario using datasets in multiples of 50, 150, 250, and 350 voice data. The results obtained have achieved 95% accuracy on the training data of 350 English voice data. The analysis carried out is to find the best composition on the Sasak language dataset by comparing the accuracy of the test results with the accuracy of the previous training results on the English dataset. From the training and testing process that has been carried out, the results obtained show that the best dataset composition for Sasak language is with nine speakers. This illustrates that the Sasak language requires less human resources compared to the English dataset which involves more than 30 speakers in 50 words. This has a positive impact on saving resources and time required in the development of Sasak language speech recognition system.**

Key words**: Analysis, Sasak Language, CNN, Best Composition, MFCC**

I. PENDAHULUAN

Pemrosesan suara adalah aplikasi pemrosesan sinyal digital (*Digital Signal Processing*) untuk memproses dan atau menganalisis sinyal suara. Ada beberapa penelitian yang dapat dilakukan dalam bidang pengolahan suara, salah satunya adalah *speech recognition* [1]. Seiring dengan berkembangnya waktu, *speech recognition* telah berkembang dengan dukungan teknologi mutakhir dan memungkinkan kemampuan pengenalan suara yang lebih canggih. Pada saat ini, pemrosesan suara digunakan untuk menggantikan peran *input keyboard* dan *mouse*. Seperti yang diterapkan pada fitur Cortana di laptop, layaknya asisten pribadi yang dapat kita perintahkan untuk melakukan pencarian tertentu dan lainnya. Cortana akan menanggapi suara yang diterima dan merespon sesuai dengan suara yang dikenali.

Penelitian sebelumnya mengenai konversi *speech-to-text* menggunakan metode *Convolutional Neural Network* (CNN) pernah dilakukan untuk klasifikasi sentimen ulasan film Indonesia. Pada penelitian ini data yang digunakan adalah audio dari video *review* 5 film Indonesia yang ada di Youtube. Berdasarkan pengujian yang telah dilakukan, metode ini menghasilkan rata-rata nilai AUC dan akurasi dari data *testing* berturut-turut sebesar 0,845 dan 0,837 [2].

Dengan pertumbuhan pariwisata internasional yang semakin membaik sejak diadakannya kegiatan motoGP di pulau Lombok, menyebabkan banyaknya turis lokal bahkan mancanegara yang datang mengunjungi *event* dan tentunya akan bersosialisasi dengan warga setempat. Pemerintah daerah sudah mengadakan pelatihan Bahasa Inggris bagi warga lokal [3]. Namun dengan keterbatasan masyarakat dalam penggunaan Bahasa Inggris dan kurangnya penutur bahasa Sasak dalam menerjemahkan percakapan dapat menjadi tantangan tersendiri bagi mereka. Pada penelitian ini, penulis akan melakukan analisis kebutuhan dataset pada algoritma *speech to text* bahasa Sasak dan Bahasa Inggris. Dengan menggunakan metode *Convolutional Neural Network* (CNN) yang cocok untuk pengolahan dataset dalam jumlah cukup banyak dan dapat melakukan pengenalan suara dengan akurasi yang cukup tinggi.

Analisis dataset merupakan cara dalam memahami, memproses, dan mengeksplor dataset untuk mengenali pola, korelasi, dan informasi yang terdapat di dalamnya. Dengan melakukan analisis dataset dapat membantu dalam pengambilan keputusan berdasarkan data yang ada. Analisis ini dilakukan dengan menggunakan perbandingan dataset Bahasa Inggris karena keterbatasan data suara bahasa Sasak dan Bahasa Inggris merupakan bahasa yang penggunaannya banyak di internet dalam bentuk suara (dataset tersedia dalam jumlah besar) dan lebih mudah dicari. Bahasa Inggris juga merupakan salah satu bahasa

yang paling banyak diteliti dalam konteks pengenalan suara [4]. Dengan dataset Bahasa Inggris, dapat digunakan sebagai pemodelan awal untuk melatih model dataset yang lebih kecil seperti suara bahasa Sasak. Model yang dihasilkan digunakan sebagai tolak ukur untuk melihat sejauh mana performa model yang dibangun dan membandingkan dengan hasil pengenalan bahasa Sasak.

II. TINJAUAN PUSTAKA

Terdapat beberapa penelitian terkait dengan *speech to text* menggunakan metode CNN. Penelitian pertama berjudul "*Speech to text Conversion using Deep Learning Neural Net Methods*" [5]. Penelitian ini dilakukan untuk meneliti dan mengevaluasi berbagai metode yang digunakan dalam konversi STT, dan menemukan metode paling efisien yang dapat disesuaikan dengan kedua proses konversi tersebut. Berdasarkan penelitian yang telah dilakukan, performa model *Deep Neural Network* dapat diandalkan untuk menguji dan memvalidasi audio dengan mencapai tingkat akurasi sebesar 95.3%.

Penelitian kedua berjudul "Implementasi Sistem Pesan Via Suara : Konversi Suara Ke Teks Pada Aplikasi Pengiriman Pesan Berbahasa Indonesia" [6]. Pada penelitian ini menggunakan data rekaman suara yang diucapkan dalam Bahasa Indonesia. Penelitian ini menghasilkan tingkat akurasi tertinggi secara *real time* pada *dependent speech* dengan persentase 17,78% dengan jumlah tiga data latih. Sedangkan pada *independent speech* didapat akurasi tertinggi 11,11% dengan jumlah satu data latih.

Penelitian lainnya dengan judul "*Speech-To-Text Conversion (STT) System Using Hidden Markov Model (HMM)*". Penelitian ini menggunakan 50 *file* audio yang telah direkam. Dari pengujian yang telah dilakukan, didapatkan hasil rata-rata tingkat pengenalan sebesar 87,6% [7].

Penelitian keempat berjudul "*Deep Learning Based Bangla Speech-to-Text Conversion*" [8]. Penelitian ini menggunakan 1400 data suara dengan format wav dan jumlah total 170 kalimat. Data tersebut dilatih menggunakan *deep recurrent neural networks* dan menghasilkan akurasi lebih dari 95% untuk data pelatihan dan akurasi 50% untuk pengujian.

Penelitian kelima berjudul "*Multilingual Speech to Text using Deep Learning based on MFCC Features*" [9]. Penelitian ini menggunakan 1920 rekaman dan dilatih menggunakan MFCC yang diekstrak. Saat diuji dapat memberikan akurasi 85% dan ketika diuji langsung oleh pengguna, itu memberikan akurasi 71%.

A. Bahasa Sasak

Bahasa Sasak adalah bahasa yang dituturkan oleh suku Sasak di Pulau Lombok. Pada data tahun 2010 diperkirakan bahwa penutur asli bahasa ini sekitar 2,7 juta. Bahasa Sasak lebih banyak digunakan sebagai bahasa lisan dibandingkan tulisan. Terdapat lima dialek bahasa Sasak yaitu dialek kuto-kutè (utara), nggeto-nggetè (tenggara), menomenè (tengah), ngeno-ngenè (tengah timur, tengah barat) dan meriaq-meriku (tengah selatan) [3].

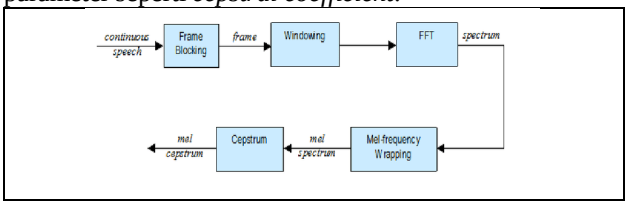
B. Suara Digital

Sinyal digital merupakan hasil pengolahan data yang dapat mengubah suara gelombang analog yang dihasilkan oleh manusia atau benda lainnya menjadi urutan bilangan 0 dan 1 sehingga dapat dapat dikenali oleh perangkat digital [1].

1. *Analog to Digital Conversion* (ADC)
   Analog To Digital Conversion (ADC) merupakan komponen elektronik yang mengkonversi amplitudo gelombang suara menjadi representasi digital dari suara [10].
2. File WAV
   WAV (*Waveform Audio File Format*) adalah format standar audio dari Microsoft dan IBM. WAV adalah data audio yang tidak dikompresi, yang berarti data audio disimpan semuanya secara langsung di *harddisk* tanpa pengurangan kualitas [10].
3. File MP3
   MP3 adalah singkatan dari MPEG (*Moving Picture Expert Group*)-1 *audio layer* III merupakan format audio yang digunakan untuk kompresi data suara dan dapat menghilangkan informasi yang tidak terdengar oleh telinga manusia [10].

C. Mel Frequency Cepstrum Coefficient (MFCC)

Mel Frequency Cepstrum Coefficient (MFCC) adalah ekstraksi yang sering diterapkan di bidang *speaker recognition* dan *speech recognition*. Karakteristik yang disebut MFCC menentukan bagaimana parameter *cepstral coeficient* akan terlihat. Dengan merepresentasikan file audio, ekstraksi dari *Mel Frequency Cepstrum Coefficient* (MFCC) dapat mengubah gelombang suara menjadi parameter seperti *cepstral coefficient*.



Gambar 1. Diagram Blok MFCC

Berdasarkan Gambar 1. setiap tahapan memiliki fungsi tertentu, *Frame Blocking* memisahkan sinyal audio menjadi beberapa *frame*, *windowing* untuk meruncingkan *frame* di awal dan di akhir, FFT untuk mengubah *frame* dari domain waktu ke dalam domain frekuensi, *mel frequency wrapping* untuk menghasilkan nilai *mel spectrum*, dan *cepstrum* untuk mengubah *spectrum log mel* menjadi domain waktu [11].

D. Speech Recognition

Speech recognition atau pengenalan suara merupakan teknologi yang menggunakan suara sebagai *input* ke dalam sistem [2]. Pengenalan suara menggunakan suara atau ucapan sebagai *input* untuk menjalankan program komputer yang kemudian dibandingkan dengan *database*

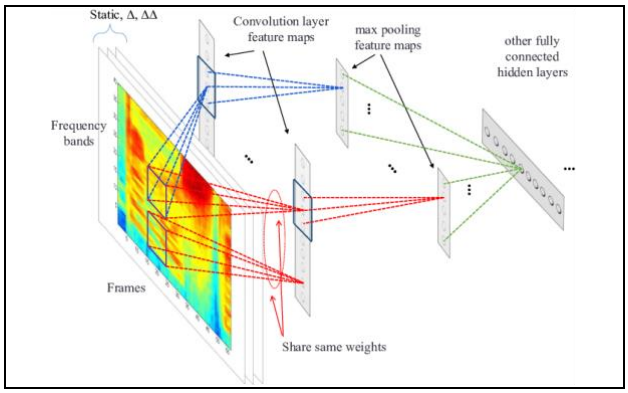
yang ada. Konverter akan mengubah suara ini dari suara fisik menjadi sinyal listrik, yang kemudian akan diubah kembali menjadi data digital. Setelah itu, dapat digunakan untuk menyalin transkripsi audio ke teks. Proses ini biasa disebut *speech-to-text* (STT).

E. Jaringan Saraf Tiruan

Jaringan saraf tiruan atau *artificial neural network* merupakan jaringan neuron yang rumit di otak manusia, berfungsi sebagai inspirasi untuk jaringan saraf tiruan, yang merupakan jaringan komputasi. Sebuah jaringan saraf khas memiliki banyak neuron buatan, atau unit, yang disusun dalam *layer input*, *hidden layer* atau *layer* tersembunyi, dan *layer output* [12]. Jaringan saraf tiruan mampu mengenali tindakan berbasis data yang direkam sebelumnya. Data sebelumnya akan dipelajari oleh jaringan saraf tiruan sehingga mampu untuk mengambil keputusan terhadap data yang belum pernah dipelajari.

F. Convolutional Neural Network

Convolutional Neural Networks atau jaringan saraf *convolutional* (CNN) adalah jaringan saraf tiruan *feed-forward* yang dalam, yang mempertahankan struktur hierarkis melalui studi representasi fitur internal dan generalisasi fitur dalam masalah gambar umum seperti pengenalan objek dan masalah penglihatan komputer lainnya. Ini tidak hanya terbatas pada gambar, ini juga dapat menghasilkan terobosan untuk masalah dengan pengenalan suara dan pemrosesan bahasa alami [12]. Pada Gambar 2. terlihat *Convolutional Neural Network* mempunyai enam lapisan, yaitu: *convolutional layer*, *pooling layer*, *normalization layer*, *Rectified Linear Units layer*, *fully connected layer*, dan *loss layer* [13].



Gambar 2. Struktur CNN Untuk Pengenalan Suara

1. *Convolutional Layer*
   Lapisan dasar CNN adalah *convolutional layer*. Pada lapisan ini, filter yang hanya tumpang tindih sebagian akan memindai semua bidang reseptif. Setiap neuron berbagi berat koneksi (*weight sharing*) sebagai hasil pemindaian untuk filter tumpang tindih parsial.
2. *Pooling Layer*
   Output dari setiap *cluster* neuron pada lokasi kernel yang sama disimpulkan oleh *pooling layer*. *Pooling layer* digunakan untuk menjaga ukuran data tetap konsisten. Dengan *layer* ini, representasi data menjadi lebih kecil dan lebih mudah diatur, serta lebih mudah untuk mengontrol *overfitting.*
3. *Normalization Layer*
   Gambar *input* dinormalisasi menggunakan *normalization layer*. Perbedaan signifikan dalam rentang nilai diatasi dengan normalisasi gambar *input.*
4. *Recitified Linear Units (ReLU) Layer*
   Bidang reseptif dari lapisan konvolusional tidak terpengaruh oleh lapisan *Rectified Linear Units (ReLU) Layer*, yang digunakan untuk meningkatkan jumlah nonlinier fungsi keputusan dan jaringan secara keseluruhan.
5. *Fully Connected Layer*
   Lapisan seperti *multilayer perceptron* (MLP) adalah *Fully Connected Layer*. Perkalian matriks digunakan dalam lapisan ini, diikuti oleh bias offset.
6. *Loss Layer*
   Lapisan *output* CNN disebut *Loss Layer*. Jika ada dua kelas klasifikasi untuk gambar, *sigmoid loss* adalah *loss layer* yang digunakan, dan *output* dari lapisan ini mengikuti distribusi Bernoulli.

G. Anaconda

Python memiliki aplikasi bernama Anaconda yang tersedia sebagai perangkat lunak *open source* untuk bahasa pemrograman *Python* dan R. Semua bentuk komputasi ilmiah, termasuk pembelajaran mesin, pemrosesan data besar, analitik prediktif, dan banyak lagi, sangat bergantung pada *Python*. Dengan menambahkan Anaconda ke sistem operasi, sudah dapat mengakses sejumlah *package* yang biasanya siap untuk digunakan langsung. Pengguna dapat menggunakan berbagai *package*, termasuk *Numpy*, *Pandas*, dan *Scikit-Learn*. Kemampuan pengguna untuk memilih versi bahasa pemrograman *Python* yang ingin mereka gunakan adalah salah satu manfaat lainnya [14].

H. Python

Python merupakan salah satu bahasa pemrograman paling populer yang sering dimanfaatkan dalam berbagai aplikasi seperti ilmu data, pembelajaran mesin, dan pengujian perangkat lunak. Karena *Python* sebelumnya telah disebut sebagai bahasa pemrograman tingkat lanjut atau tingkat tinggi, sintaksnya mudah dipahami. *Python* adalah bahasa pemrograman yang banyak digunakan dan disukai, terutama saat membuat program untuk pembelajaran mesin. Ini agar program komputasi yang kompleks dapat dirancang menggunakan banyak fungsi *library Python*. *Library* populer termasuk *Scikit-learn*, *NumPy*, dan *Matplotlib* [14].

I. Artificial Intelligent

Artificial Intelligent atau dikenal juga sebagai kecerdasan entitas ilmiah, adalah kecerdasan yang ditambahkan ke sistem yang dapat diatur dalam konteks ilmiah. Andreas Kaplan dan Michael Haenlein mendefinisikan kecerdasan buatan sebagai kapasitas sistem untuk secara akurat menafsirkan data eksternal, untuk belajar dari data tersebut dan untuk menerapkan pembelajaran tersebut untuk mencapai tujuan dan tugas

tertentu melalui kemampuan beradaptasi. AI dapat meniru perilaku manusia dengan bertindak dengan cara yang dapat dicirikan sebagai pintar atau inventif. Ada beberapa cabang ilmu yang mengaplikasikan kecerdasan buatan seperti sistem pakar, *game*, *fuzzy logic*, *artificial neural network*, dan robot [15].

J. Tensorflow

Library open source yang disebut Tensorflow digunakan dalam proyek pembelajaran mesin skala besar dan komputasi numerik. Tim Google Brain mengembangkan tensorflow yang menggabungkan berbagai model dan algoritma *machine learning*, termasuk *deep learning* (*neural network*). Tensorflow memungkinkan pelatihan dan pengoperasian model yang lebih cepat sekaligus memberi fleksibilitas untuk iterasi dengan cepat [12].

K. Keras

Keras merupakan sebuah *framework* yang dikembangkan untuk membantu pembelajaran terhadap komputer. Salah satu library dalam jaringan saraf tiruan tingkat tinggi yang ditulis dengan *Python* dan mampu berjalan di Tensorflow, CNTK, atau Theano disebut Keras. *Library* ini menawarkan fitur yang dapat digunakan untuk berkonsentrasi pada pengembangan tentang *deep learning* [16].

III. METODE PENELITIAN

A. Alat dan Bahan

Alat yang digunakan terdiri dari perangkat keras dan perangkat lunak. Sementara bahan yang digunakan terdiri dari data yang dibutuhkan berupa rekaman suara. Alat dan bahan tersebut dapat dilihat pada Tabel I dan Tabel II.

1. Perangkat Keras

TABEL I. PERANGKAT KERAS

No	Perangkat Keras	Spesifikasi
1	Laptop	Lenovo Thinkpad dengan processor Intel Core i5, RAM 16 GB

2. Perangkat Lunak

TABEL 1. PERANGKAT LUNAK

No	Perangkat Lunak	Spesifikasi
1	Sistem Operasi	Windows 10 64-bit
2	Bahasa Pemrograman	*Python*
3	Lingkungan Pemrograman	Pycharm

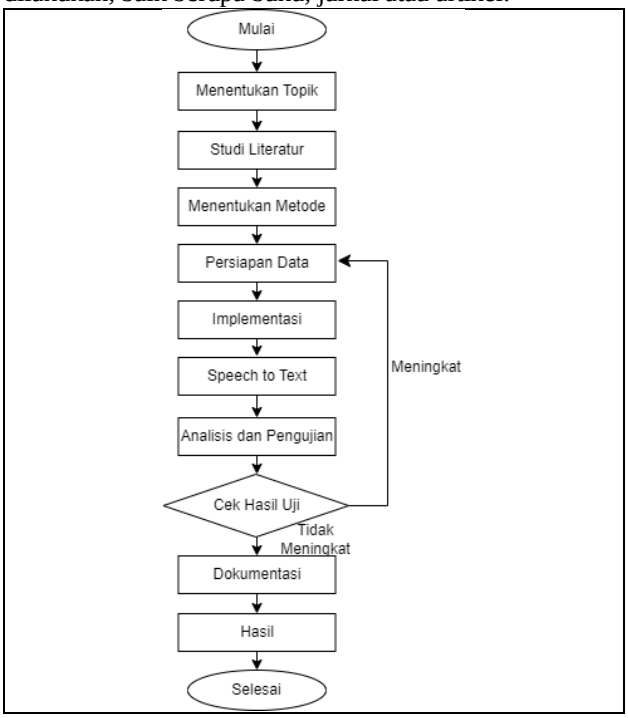
3. Bahan Penelitian

Sumber data yang digunakan sebagai bahan pada penelitian ini ini berupa rekaman suara bahasa Sasak dari masyarakat asli dan data suara Bahasa Inggris. Proses perekaman suara bahasa Sasak dilakukan dengan menggunakan *smartphone* dan aplikasi *voice recorder pro* dari masing-masing responden. Sedangkan untuk data suara Bahasa Inggris diunduh pada https://ai.googleblog.com/2017/08/launching-speech-commands-dataset.html. Terdapat 10 kata dalam bahasa Sasak dan Bahasa Inggris, yang direkam dengan suara bahasa Sasak kurang lebih 3500 data dan suara Bahasa Inggris sebanyak 17400 data.

Gambar 3. Dataset Suara Bahasa Sasak

B. Alur Penelitian

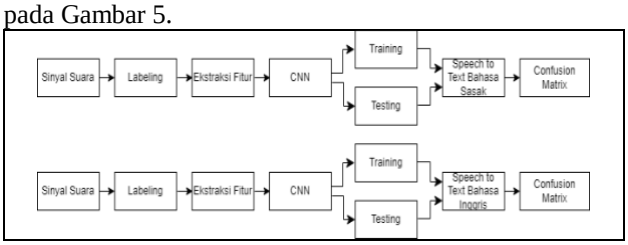
Secara umum alur penelitian pada penelitian ini dapat digambarkan seperti diagram alir pada Gambar 4. Sebelumnya telah dilakukan studi literatur untuk menambah pengetahuan dalam memahami apa yang akan dilakukan. Studi literatur dilakukan dengan membaca berbagai referensi terkait penelitian serupa yang akan dilakukan, baik berupa buku, jurnal atau artikel.



Gambar 4. Diagram Alir Alur Penelitian

C. Perancangan Sistem

Secara umum perancangan sistem dari penelitian ini yang akan dilakukan melalui beberapa tahap hingga mendapatkan hasil akhir yang ingin dicapai dapat dilihat pada Gambar 5.

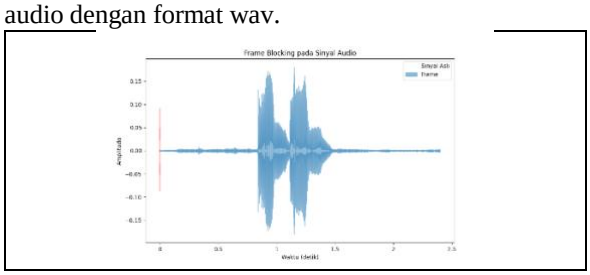


Gambar 5. Diagram Blok Sistem

Proses perancangan sistem dalam penelitian ini melalui pendekatan yang sistematis untuk mencapai hasil yang optimal dalam pelatihan dan pengujian dataset. Proses ini melakukan beberapa tahapan yaitu pelatihan dimulai dengan 50 data suara Bahasa Inggris dengan hasil akurasi yang diperoleh misalnya sebesar 70%. Langkah selanjutnya jumlah dataset ditambah menjadi 150 data suara dan mendapat akurasi 80%. Kemudian pada pelatihan ketiga dengan 250 data suara dan akurasi meningkat menjadi 85%. Selama proses penambahan dataset dan akurasi yang dihasilkan masih menunjukkan peningkatan yang signifikan, maka pelatihan akan terus berlanjut hingga akurasi yang diperoleh konsisten atau maksimal data suara yang digunakan mencapai 1750 data Bahasa Inggris. Setelah akurasi yang diperoleh optimal dan konsisten, kemudian dilakukan pengujian dengan menggunakan dataset bahasa Sasak. Hasil dari pelatihan Bahasa Inggris yang telah dilakukan sebelumnya akan menjadi acuan untuk mengukur kualitas pengenalan suara bahasa Sasak.

1. Sinyal Suara

   Pada penelitian ini, sinyal suara yang digunakan adalah rekaman suara bahasa Sasak dan Bahasa Inggris yang terdiri dari masing-masing 10 kata dengan jumlah 250 rekaman suara bahasa Sasak untuk tiap kata dan 1740 rekaman untuk tiap kata dalam suara Bahasa Inggris. Sinyal suara yang digunakan adalah dalam bentuk file audio dengan format wav.



Gambar 6. Sinyal Suara Kata Adeng

2. *Labeling*

   Tahap ini dilakukan untuk memberi label data suara yang dimasukkan dan dikelompokkan sehingga memudahkan dalam proses selanjutnya. Proses ini bertujuan untuk menganalisis atau pengaplikasian dalam model *machine learning. Labeling* dapat diperoleh dari hasil ekstraksi fitur yang telah dilakukan. Hasil *labelimg* dapat dilihat pada Tabel III.

TABEL III. LABEL DATASET

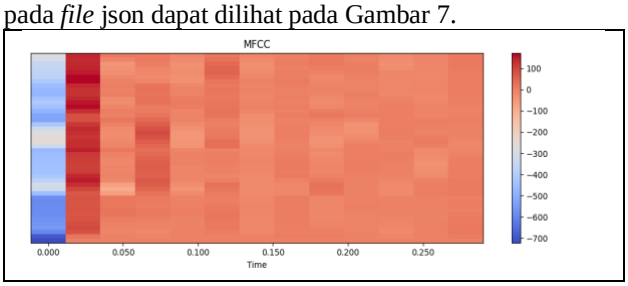
Bahasa Inggris	Label	Bahasa Sasak	Label
Bed	0	Adeng	0
Bird	1	Aiq	1
Cat	2	Anjah	2
Dog	3	Bansu	3
Five	4	Bareh	4
Go	5	Berajah	5
Happy	6	Endeqman	6
Nine	7	Gaweq	7
No	8	Inaq	8
On	9	Lekak	9

D. Metode yang Diajukan

1. Ekstraksi Fitur MFCC

Metode ekstraksi MFCC telah sering diterapkan untuk penelitian terkait suara manusia. Untuk mengekstrak fitur dari *file* audio, dapat menggunakan MFCC. Tahap ekstraksi fitur MFCC memerlukan sejumlah parameter, termasuk batas waktu 60 detik, panjang *frame* 40, dan tumpang tindih (*overlap*) 40%. Ini juga menggunakan *hamming window*, yang diubah menjadi *windowing* di salah satu tahap MFCC. Langkah selanjutnya adalah memilih data sebagai salah satu skenario pengujian setelah koefisien MFCC diperoleh [17].

Ekstraksi fitur MFCC merupakan tahap awal dari rangkaian proses yang ada setelah semua data dikumpulkan. Melalui ekstraksi fitur ini diperoleh sebuah *file* json yang di dalamnya terdapat beberapa informasi dalam bentuk matriks mengenai data suara yang akan dilatih dan diuji. Beberapa informasi tersebut seperti *mapping*, *labels*, *MFCCs*, dan *files*. *Mapping* berisi daftar kategori pengelompokkan audio berdasarkan kata yang dikenali. *Mapping* sering digunakan untuk menghubungkan label numerik pada matriks *labels* dengan label kategori. *MFCCs* berisi data *array* atau daftar numerik *frame* waktu dalam suara di mana pada penelitian ini jumlah koefisien sebanyak 13. Kemudian untuk *files* memuat informasi mengenai *path* audio yang sesuai dengan dataset yang digunakan. Koefisien yang terdapat pada *file* json dapat dilihat pada Gambar 7.

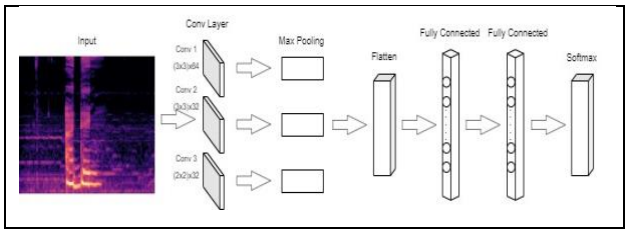


Gambar 7. Plotting MFCC

2. CNN

Ada beberapa lapisan dari struktur CNN seperti *convolution layer*, *max pooling*, dan *fully connected layer*. CNN hanya dapat digunakan pada data yang memiliki struktur dua dimensi, seperti gambar dan suara, karena sifat proses konvolusi [18]. Model yang akan dibangun menggunakan metode CNN dengan *library* Keras yang diimplementasikan dalam TensorFlow. Model ini berupa *layers* atau lapisan yang terdiri dari 3 lapisan konvolusi dengan lapisan konvolusi pertama memiliki 64 kernel dengan ukuran 3x3 dan menggunakan aktivasi ReLU. Setelah konvolusi dilakukan selanjutnya operasi *max-pooling* 2D dengan ukuran (3,3). Pada lapisan konvolusi kedua sama dengan lapisan pertama, namun pada lapisan ini menggunakan 32 kernel. Selanjutnya pada lapisan konvolusi terakhir memiliki ukuran kernel (2,2) dan ketiga lapisan menggunakan normalisasi *batch* dan *max-pooling*. Kemudian *output* dari lapisan konvolusi akan diambil oleh *flatten layer* dan diubah ke dalam bentuk vektor satu dimensi untuk dimasukkan ke lapisan *dense*. Terdapat 2

lapisan *dense* dengan lapisan pertama memiliki 64 *neuron* dengan aktivasi ReLU dan *dropout* sebesar 0,3 yang berarti 30% dari *neuron* akan dimatikan secara acak selama pelatihan atau *training*. Kemudian pada lapisan kedua yaitu lapisan *output* dengan 10 *neuron* (sesuai dengan jumlah kelas yang akan diprediksi) dan menggunakan aktivasi *softmax* untuk memprediksi kelas. Pada model ini digunakan *optimizer* Adam dengan *learning rate* sebesar 0,0001. Adam merupakan salah satu *optimizer* yang banyak digunakan dan efektif dalam pembelajaran mesin. Arsitektur dari model CNN dapat dilihat pada Gambar 8 di bawah ini.



Gambar 8. Model CNN

E. Cara Analisis

Pada penelitian ini dilakukan tahap *training* dan *testing*. Dataset yang telah dijadikan MFCC akan diproses sebagai data *training* menggunakan CNN untuk ditampilkan prosesnya dan menampilkan akurasi. Dari hasil *training*, kemudian diproses untuk diklasifikasi sehingga menampilkan hasil dalam bentuk teks.

1. Persiapan Data
   Data suara harus diproses terlebih dahulu menggunakan MFCC yang melibatkan penerapan filter pada sinyal suara. Hasil yang didapatkan adalah serangkaian vektor numerik yang merepresentasikan sinyal suara.
2. *Training*
   Setelah data suara dipersiapkan, langkah selanjutnya adalah membuat model CNN. Model ini terdiri dari beberapa lapisan konvolusi dan pengurangan dimensi, diikuti oleh beberapa lapisan *dense* (seperti pada model *neural network* biasa). Model CNN pada pemrosesan suara dapat menggunakan filter 1D. Setelah model CNN dibuat, langkah selanjutnya adalah melatihnya pada data pelatihan. Selama pelatihan, model akan menerima serangkaian vektor numerik MFCC dan mencoba mempelajari pola dan fitur di dalamnya. Model akan diberi label untuk menunjukkan kelas yang benar, dan model akan menyesuaikan parameter internalnya untuk meningkatkan akurasi pada data pelatihan. Setelah model CNN dilatih pada data pelatihan, langkah selanjutnya adalah memvalidasi kinerjanya pada data validasi. Data validasi adalah data yang tidak pernah dilihat oleh model selama pelatihan. Validasi membantu memastikan bahwa model tidak hanya mempelajari data pelatihan dengan baik, tetapi juga mampu menggeneralisasi pola dan fitur pada data yang tidak dilihat sebelumnya.
3. Pengujian
   Proses pengujian mirip dengan proses pelatihan, tetapi data yang digunakan untuk pengujian adalah data yang tidak dilihat oleh model sebelumnya. Setelah proses pengujian selesai, langkah terakhir adalah mengevaluasi kinerjanya pada data pengujian. Data pengujian adalah data yang sepenuhnya tidak terlihat oleh model sebelumnya dan digunakan untuk mengetahui apakah model mampu memprediksi label dengan benar pada data baru. Kinerja model dapat dievaluasi dengan menggunakan beberapa metrik, seperti akurasi, presisi, *recall*, dan *F1-score*, hal ini nantinya akan dilakukan berulang dengan kelipatan 100 data mulai dari 50, 150, 250, 350, 450 sampai maksimum 1750 atau sampai data menunjukkan kondisi stabil. Metrik-metrik ini memberikan informasi tentang seberapa baik model dapat memprediksi label yang benar pada data pengujian.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3\text{-}1)$$

$$Precision = \frac{TP}{TP+FP} \quad (3\text{-}2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3\text{-}3)$$

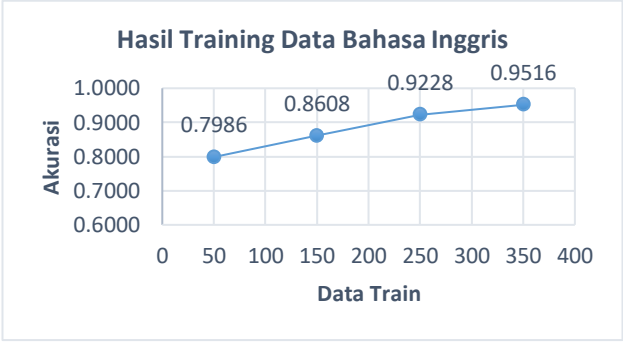
$$F-score = 2 . \frac{precision \, . \, recall}{precision + recall} \quad (3\text{-}4)$$

Dimana:
TP = True Positive (Suara benar dianggap benar)
TN = True Negative (Suara benar dianggap salah)
FP = False Positive (Suara salah dianggap benar)
FN = False Negative (Suara salah dianggap salah)

IV. HASIL DAN PEMBAHASAN

A. Training Data

Pada proses *training*, dataset yang digunakan adalah data Bahasa Inggris yaitu 10 kata yang diambil secara acak dari keseluruhan kata yang tersedia, kemudian diambil 50 rekaman setiap kata untuk dilakukan *training* pertama kali. Selanjutnya ditambah 100 rekaman atau kelipatan 100 untuk *training* berikutnya hingga mendapat nilai akurasi yang stabil yaitu sekitar lebih dari 92% dengan jumlah data *train* sebanyak 350 data rekaman. Hasil *training* dataset Bahasa Inggris dapat dilihat pada Gambar 9.

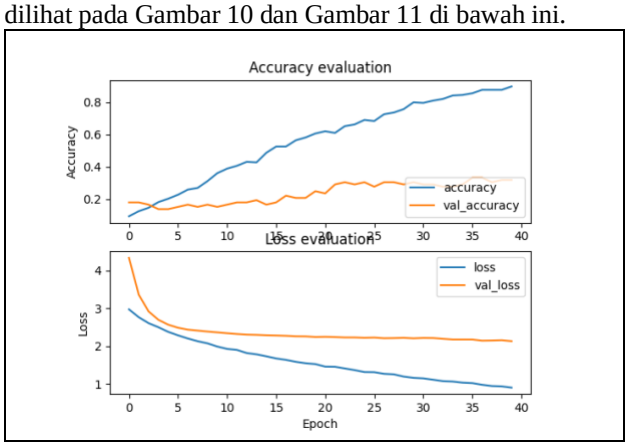


Gambar 9. Hasil Training Data Bahasa Inggris

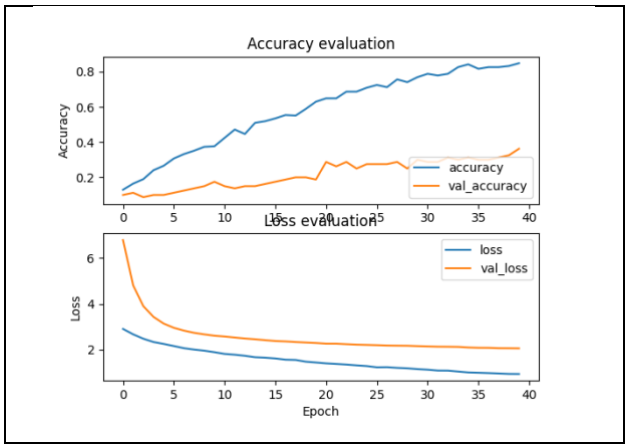
B. Pengujian

Pada tahap pengujian yang dilakukan dengan menggunakan dataset bahasa Sasak, hasil pengujian mengacu pada hasil pelatihan sebelumnya yang dilakukan pada dataset Bahasa Inggris. Data latih Bahasa Inggris menghasilkan akurasi yang secara konsisten meningkat, sehingga dengan penambahan data maka tingkat akurasi

juga meningkat hingga mencapai angka lebih dari 90. Akurasi ini dianggap sebagai titik stabil dalam pelatihan. Hasil akurasi sebagai perbandingan dari 50 data pelatihan dalam Bahasa Inggris dan 50 data bahasa Sasak dapat dilihat pada Gambar 10 dan Gambar 11 di bawah ini.



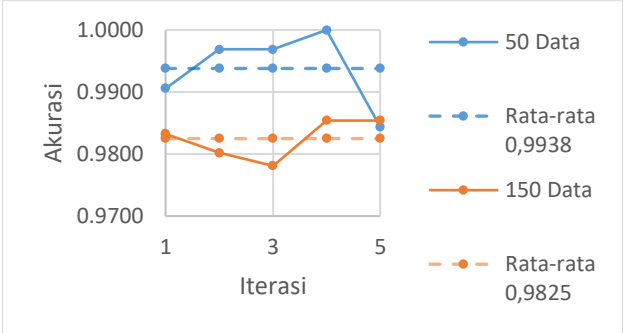
Gambar 10. Hasil Pelatihan 50 Data Bahasa Inggris



Gambar 11. Hasil Pelatihan 50 Data Bahasa Sasak

C. Analisis

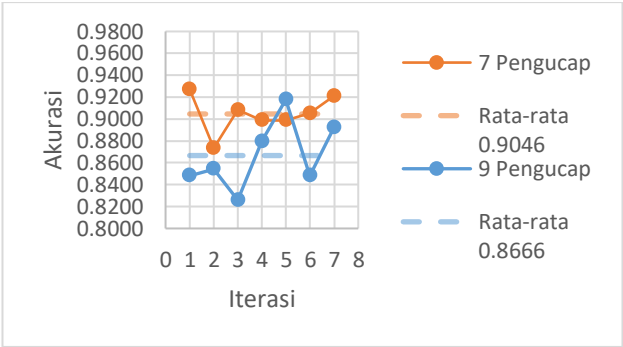
Analisis dilakukan untuk mencari komposisi terbaik dari data bahasa Sasak dengan perbandingan dari akurasi pelatihan Bahasa Inggris. Peneliti telah melakukan beberapa kali percobaan dalam pengujian data hingga mencapai komposisi atau jumlah variasi suara yang sesuai untuk menghindari terjadinya *overfitting* namun dengan meminimalisir jumlah variasi tersebut. Pada percobaan pertama hanya menggunakan satu suara pengucap atau individu pada 50 data kemudian menambah satu suara yang berbeda lagi pada pengujian 150 data. Hasil yang diperoleh yaitu pengujian pada 50 data mendapatkan akurasi yang lebih tinggi dari pengujian 150 data atau bisa dikatakan akurasi semakin menurun walaupun tidak begitu signifikan jika data terus ditambah. Hal ini disebut dengan *overfitting*, sehingga proses pengujian dihentikan dan menambah beberapa pengucap lagi pada dataset. Hasil pengujian dengan suara tanpa adanya variasi dapat dilihat pada Gambar 12.



Gambar 12. Hasil Testing Overfitting

Dari hasil tersebut, penguji melakukan penambahan jumlah pengucap pada dataset bahasa Sasak untuk menghilangkan *overfitting*. Proses penambahan ini dilakukan secara bertahap dengan memperhatikan tingkat akurasi yang dihasilkan pada setiap pengujian dengan jumlah pengucap yang berbeda. Pertama, dataset yang terdiri dari 50 data dengan hanya satu orang pengucap ditambah menjadi dua orang. Namun hasilnya masih menunjukkan kecenderungan *overfitting*. Kemudian dataset diperluas lagi dengan menambahkan pengucap ketiga. Akan tetapi hasilnya tetap menunjukkan ciri-ciri *overfitting* yang membuktikan bahwa perbedaan antara ketiga pengucap masih belum cukup untuk membuat model melakukan generalisasi dengan baik.

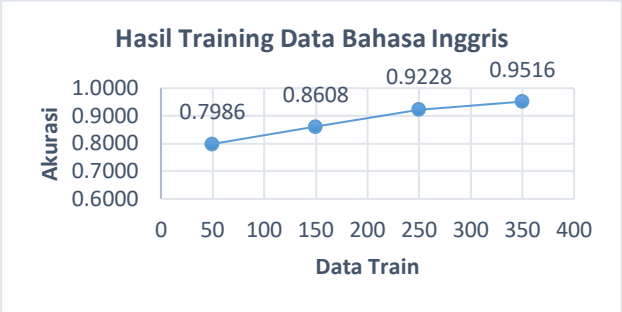
Proses dilanjutkan dengan penambahan lebih banyak pengucap. Dataset diperluas menjadi tujuh pengucap dan terakhir sembilan pengucap dalam 50 data bahasa Sasak. Hasil yang diperoleh adalah tingkat akurasi tidak berbeda secara signifikan dari hasil pelatihan sebelumnya pada dataset Bahasa Inggris dengan jumlah pengucap lebih dari 30 orang. Ini menunjukkan bahwa model telah mampu melakukan generalisasi dengan baik. Hasil akurasi dengan tujuh pengucap dan sembilan pengucap dapat dilihat pada Gambar 13.



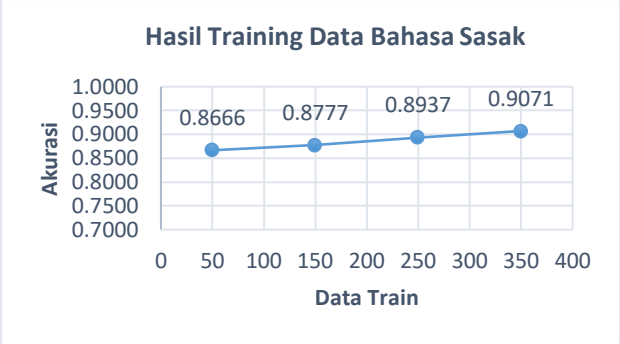
Gambar 13. Perbandingan Hasil 7 Pengucap dan 9 Pengucap

Berdasarkan pengujian yang telah dilakukan sebagai upaya untuk melihat stabilitas dan hasil yang konsisten, iterasi dilakukan sebanyak tujuh kali. Setiap iterasi merupakan langkah penting untuk memastikan bahwa hasil pengujian tidak hanya bergantung pada satu kali percobaan. Pada Gambar 4.8 terlihat bahwa tingkat akurasi meningkat dengan cukup baik mencapai angka di atas 90. Namun setelah mencapai puncaknya, akurasi mengalami penurunan dan stabil di angka 80-an.

Oleh karena itu, untuk melihat hasil yang konsisten diambil nilai rata-rata dari hasil kedua pengujian yang telah dilakukan. Nilai rata-rata akan menjadi acuan untuk menentukan jumlah pengucap yang paling cocok dan konsisten untuk dilakukan pada pengujian berikutnya yaitu 150 data, 250 data, dan terakhir 350 data. Rata-rata yang diperoleh dari kedua pengujian tersebut yaitu sebesar 0,9046 untuk data tujuh pengucap dan 0,8666 untuk data sembilan pengucap. Hasil yang paling mendekati dengan hasil pelatihan Bahasa Inggris adalah data dengan sembilan pengucap dan menjadi pilihan yang dianggap optimal untuk dilanjutkan ke pengujian selanjutnya.



Gambar 14. Hasil Training Bahasa Inggris



Gambar 15. Hasil Training Bahasa Sasak

Pada Gambar 15 menunjukkan hasil akurasi dari pengujian bahasa Sasak sampai dengan penggunaan dataset sebanyak 350 data. Dapat dilihat bahwa hasil pengujian telah mencapai tingkat akurasi yang diinginkan. Hasil ini sangat sejalan dengan hasil pelatihan yang telah dilakukan sebelumnya pada dataset Bahasa Inggris.

Dari proses pelatihan dan pengujian yang telah dilakukan, dapat disimpulkan bahwa komposisi dataset terbaik untuk bahasa Sasak dengan hasil akurasi yang optimal dan stabil adalah menggunakan sembilan pengucap. Hal ini menunjukkan bahwa dataset bahasa Sasak memerlukan jumlah sumber daya manusia yang jauh lebih sedikit dibandingkan dengan dataset Bahasa inggris yang menggunakan lebih dari 30 pengucap pada 50 data. Dengan mendapat komposisi yang sesuai, penelitian ini menunjukkan bahwa bahasa Sasak berhasil diuji dan memperoleh hasil akurasi yang tinggi bahkan dengan menggunakan dataset yang lebih sedikit. Ini berdampak positif pada penghematan sumber daya dan waktu dalam pengembangan sistem pengenalan suara bahasa Sasak.

V. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan penelitian yang telah dilakukan, maka diperoleh kesimpulan sebagai berikut:

1) Implementasi algoritma konversi suara bahasa Sasak dan Bahasa Inggris menjadi teks berhasil diimplementasikan dengan menggunakan metode CNN. Dengan melalui beberapa tahap seperti pengumpulan data, *preprocessing*, *training*, validasi dan evaluasi, dan terakhir menganalisis hasil.
2) Metode CNN pada penelitian ini menggunakan tiga lapisan konvolusi di mana setiap lapisannya diikuti oleh operasi *max-pooling* 2D. Hasil dari lapisan ini diambil dan diubah dalam bentuk vektor satu dimensi dengan lapisan *flatten*. Selanjutnya vektor melawati dua lapisan *dense* yang memiliki fungsi aktivasi ReLU dan *dropout* sebesar 30%. Pada lapisan kedua menggunakan aktivasi *softmax* untuk prediksi kelas dan menggunakan *optimizier* Adam dengan *learning rate* 0,0001.
3) Komposisi dataset terbaik yang menghasilkan akurasi optimal dan stabil dari bahasa Sasak adalah sebanyak sembilan pengucap. Hal ini menggambarkan bahwa dataset bahasa Sasak membutuhkan jumlah sumber daya yang jauh lebih sedikit dibandingkan dataset Bahasa Inggris yang menggunakan lebih dari 30 pengucap dalam 50 data.

B. Saran

Berdasarkan penelitian yang telah dilakukan, berikut beberapa saran perbaikan ataupun pengembangan yang dapat dilakukan pada penelitian selanjutnya:

1) Pada penelitian selanjutnya dapat mengubah variasi dalam arsitektur CNN seperti menambahkan lapisan konvolusi untuk melihat apakah dapat meningkatkan akurasi.
2) Memperluas pengujian dengan menggunakan lebih banyak data bahasa Sasak dan Bahasa Inggris atau bahasa lainnya.
3) Mengaplikasikan model dalam kehidupan, seperti transkripsi otomatis pada sebuah rekaman untuk melihat sejauh mana model ini dapat diimplementasikan dalam kehidupan sehari-hari.

DAFTAR PUSTAKA

[1] L. R. Rabiner and R. W. Schafer, *the essence of knowledge Introduction to Digital Speech Processing Introduction to Digital Speech Processing Foundations and Trends ® in Signal Processing Introduction to Digital Speech Processing*, vol. 1, no. 12. 2007.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] I. Ayu Shafirra N, "Klasifikasi Sentimen Ulasan Film Indonesia dengan Konversi Speech-to-Text (STT) Menggunakan CNN," *J. sains dan seni ITS*, vol. 9, no. 1, pp. 2301–9271, 2020.

[3] D. Wahyudin, "Identitas Orang Sasak: Studi Epistemologis terhadap Mekanisme Produksi Pengetahuan Masyarakat

Suku Sasak," *J. Penelit. Keislam.*, vol. 14, no. 1, pp. 52–63, 2018, doi: 10.20414/jpk.v14i1.493.

[4] N, Martin, T Basaruddin, "Dataset Suara dan Teks Berbahasa Indonesia Pada Rekaman," *Jurnal Fasilkom*, vol. 11, no. 2, pp. 61–66, 2021.

[5] B. Pandipati and R. P. Sam, "Speech to text Conversion using Deep Learning Neural Net Methods," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 05, pp. 2037–2042, 2021.

[6] G. Kristian and W. Kusuma, "Implementasi Sistem Pesan Via Suara : Konversi Suara Ke Teks Pada Aplikasi Pengiriman Pesan Berbahasa Indonesia Implementation Of Voice On Message : Speech To Text On Indonesian-Language Sended Message," *e-Proceeding of Engineering*, vol. 2, no. 1, pp. 651–657, 2015.

[7] S. M. Mon and H. M. Tun, "Speech-To-Text Conversion (STT) System Using Hidden Markov Model (HMM)," *International Journal Of Scientific & Technology Research*, vol. 4, no. 06, pp. 349–352, 2015.

[8] M. T. Tausif, S. Chowdhury, M. S. Hawlader, M. Hasanuzzaman, and H. Heickal, "Deep Learning Based Bangla Speech-to-Text Conversion," *Proc. - 5th Int. Conf. Comput. Sci. Appl. Informatics, CSII 2018*, no. July, pp. 49–54, 2018, doi: 10.1109/CSII.2018.00016.

[9] P. D. Reddy, "Multilingual Speech to Text using Deep Learning based on MFCC Features," *Mach. Learn. Appl. An Int. J.*, vol. 9, no. 02, pp. 21–30, 2022, doi: 10.5121/mlaij.2022.9202.

[10] R. Rahandi, Aditya. Dian and S. Sajadin, "Analisis dan Implementasi Kompresi File Audio Dengan Menggunakan Algoritma Run Length Encoding (RLE)," *Alkhawarizmi*, vol. 1, no. 1, 2012.

[11] Punggawa Arcapada, W. Setiawan, and I. M. Arsa Suyadnya, "Rancang Bangun Model Pengidentifikasi Suara Huruf Hijaiyah Dengan Metode Mel Frequency Cepstrum Coefficient dan Convolutional Neural Network," *J. SPEKTRUM*, vol. 8, no. 4, p. 1, 2022, doi: 10.24843/spektrum.2021.v08.i04.p1.

[12] B. Raharjo, *Deep Learning dengan Python*. 2022.

[13] M. H. Ashshiddieqy, Jondri, and A. Rizal, "Klasifikasi Suara Paru Dengan Convolutional Neural Network (CNN)," *eProceedings Eng.*, vol. 7, no. 2, pp. 8506–8512, 2020.

[14] Efanntyo and A. R. Mitra, "Perancangan Aplikasi Sistem Pengenalan Wajah Dengan Metode Convolutional Neural Network (CNN) Untuk Pencatatan Kehadiran Karyawan," *J. Instrumentasi dan Teknol. Inform.*, vol. 3, no. 1, pp. 1–11, 2021.

[15] M. Siahaan, C. H. Jasa, K. Anderson, and M. Valentino, "Penerapan Artificial Intelligence (AI) Terhadap Seorang Penyandang Disabilitas Tunanetra," *Inf. Syst. Technol.*, vol. 01, no. 02, pp. 186–193, 2020.

[16] A. Santoso and G. Ariyanto, "Implementasi Deep Learning berbasis Keras untuk Pengenalan Wajah," *Emit. J. Tek. Elektro*, vol. 18, no. 1, pp. 15–21, 2018, doi: 10.23917/emitor.v18i01.6235.

[17] F. F. Surenggana *et al.*, "Klasifikasi Mood Musik Menggunakan K-Nearest Neighbor Dengan Mel Frequency Cepstral Coefficients (Mood Music Classification using K-Nearest Neighbor with Mel Frequency Cepstral)," *Jurnal Teknologi Informasi, Komputer dan Aplikasinya (JTIKA)*, vol. 4, no. 2, pp. 263–276, 2022.

[18] S. R. Suartika E. P, I Wayan, Wijaya Arya Yudhi, "Klasifikasi Citra Menggunakan Convolutional Neural Network (Cnn) Pada Caltech 101," *J. Tek. ITS*, vol. 5, no. 1, p. 76, 2016, [Online]. Available: http://repository.its.ac.id/48842/.