# Implementasi Deep Neural Network Untuk Perintah Suara Dan Pengenalan Pembicara Pada Smart Home

1st Muhammad Bimo Wirasena
Fakultas Teknik Elektro Telkom University
*Bandung, Indonesia*
wirasena@student.telkomuniversity.ac.id

2nd Muhammad Ary Murti
Fakultas Teknik Elektro
Telkom University Bandung, Indonesia
arymurti@telkomuniversity.ac.id

3rd Junartho Halomoan
Fakultas Teknik Elektro
Telkom Uniersity Bandung, Indonesia
junartho@telkomuniversity.ac.id

**Perkembangan smart home mendorong kebutuhan sistem kendali yang tidak hanya praktis melalui perintah suara, tetapi juga aman melalui identifikasi pembicara. Pada penelitian ini dirancang dan diimplementasikan sistem *smart home* berbasis Deep Neural Network (DNN) yang menggabungkan modul pengenalan perintah suara dengan modul pengenalan pembicara (speaker recognition). Sistem direalisasikan menggunakan mikrofon sebagai sensor suara yang terhubung ke Raspberry Pi 5 untuk menjalankan proses deep learning, kemudian perintah diteruskan secara nirkabel ke ESP32 untuk mengendalikan relay (quad-channel) sehingga beban listrik (misalnya lampu) dapat dinyalakan atau dimatikan. Metode pengenalan pembicara menggunakan model gabungan CNNSpeaker dan RNN BiLSTM, di mana CNN mengekstraksi pola lokal pada domain waktu–frekuensi dari fitur akustik 105 dimensi (log Mel- spectrogram, MFCC, delta, delta-delta, RMS, dan frekuensi dasar), sedangkan BiLSTM memodelkan dinamika temporal ujaran; mekanisme attention merangkum urutan fitur menjadi embedding 256 dimensi. Hasil pengujian menunjukkan akurasi klasifikasi sekitar 91% pada 900 sampel uji dengan nilai F1-score di atas 87%. Selain itu, pengujian verifikasi dengan threshold sekitar 0,7 menunjukkan hasil pengenalan pembicara yang baik.**

**Kata kunci: *Smart home, voice command, speaker recognition*, deep neural network, CNN–BiLSTM**

## I PENDAHULUAN

Dalam Teknologi smart home terus berkembang untuk meningkatkan kenyaman, efisiensi, dan keamanan dalam kehidupan sehari-hari. Salah satu fitur penting dalam smart home modern adalah kendali perintah suara, yang memungkinkan pengguna mengontrol perangkat rumah tangga dengan mudah dan efisien melalui suara seperti penelitian yang dilakukan penelitian yang telah dilakukan oleh Nugroho et al.[1]. Dari penelitian yang dikakukan Nugroho N et al., pengendalian melalui perintah suara menjadi fitur utama karena kemudahan dan kepraktisannya. Dengan perintah suara, pengguna dapat mengontrol perangkat rumah tanpa perlu menyentuh perangkat secara fisik, menciptakan interaksi yang lebih alami dan intuitif[2].

Platform seperti Raspberry Pi, sebuah komputer mini yang hemat energi dan fleksibel, sering digunakan sebagai inti dari sistem smart home berbasis perintah suara[3]. Tran T et al. telah mencoba metode menggunakan Raspberry Pi dengan membuat kapasitas untuk menjalankan sistem operasi, framework pembelajaran mesin, dan model deep learning, Raspberry Pi dapat mengelola pemrosesan sinyal suara secara real-time. Platform ini juga dapat terhubung dengan sensor, kamera, dan aktuator yang memungkinkan pengumpulan data lingkungan, pengenalan suara, serta pengambilan keputusan secara otomatis [4]. Integrasi Raspberry Pi dengan sensor dan aktuator memungkinkan sistem untuk mengumpulkan data lingkungan, memproses informasi, dan mengambil tindakan yang sesuai berdasarkan perintah suara pengguna[3].

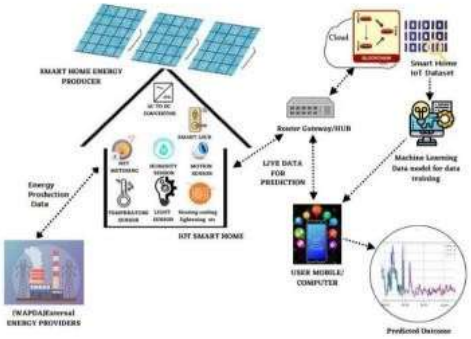
Deep Neural Network (DNN) memainkan peran kunci memiliki peran penting dalam pengenalan suara. Deep neural network dikenal memiliki kemampuan yang tinggi dalam mengenali dan mengklasifikasi pola suara secara akurat, sehingga ideal untuk mengatasi tantangan dalam pengenalan suara, termasuk variasi aksen, intonasi, dan karakteristik suara pengguna[5]. Penelitian yang telah dilakukan oleh Saxena et al bahwa implementasi DNN dalam sistem smart home memungkinkan pengenalan suara yang lebih akurat bahkan dalam kondisi lingkungan yang berisik dan kemampuan ini sangat berguna mengingat perintah suara pada smart home sering kali diberikan di lingkungan yang memiliki banyak gangguan suara latar, seperti suara televisi atau peralatan rumah tangga. Methode tersebut menggunakan algoritma Artificial Neural Network dengan akurasi 84,62%[6]. Model RNN dalam implementasi DNN dalam perintah suara mencapai tingkat akurasi tertinggi sebesar 97,3%, yang menunjukkan bahwa model ini berhasil mengklasifikasikan 97,3% dari total instance dengan benar. Selain itu, model ini juga mencapai skor presisi dan recall yang tinggi, masing-masing sebesar 97,9% dan 98,1%. Skor F1 untuk RNN juga tinggi, yaitu 97,77%. Hasil ini menunjukkan bahwa model RNN dapat mendeteksi pola stres dalam rekaman suara secara efektif dengan tingkat akurasi dan presisi yang tinggi[7]. Selain itu, penelitian P Hung et al. menyatakan penggunaan RNN dalam smart home memiliki peran penting dalam aspek keamanan. Setelah 20 percobaan pengrnalan suara, dapat dilihat bahwa model BiLSTM (98,19%) memberikan hasil yang lebih baik dibandingkan model LSTM (97,09%). Namun, terdapat masalah overfitting pada kedua model tersebut. Selanjutnya, teknik dropout regularization digunakan untuk mengurangi overfitting dan meningkatkan generalisasi teknologi jaringan saraf tiruan mendalam (Deep Neural Networks)[8].

## II KAJIAN TEORI

### *a. Smart Home*

*Smart Home* terdiri dari serangkaian sensor dan aktuator, perangkat berkemampuan IoT yang dapat dikontrol oleh

perangkat lunak melalui aplikasi atau suara[2]. Menurut Alam et al., implementasi IoT untuk rumah pintar diatur dengan arsitektur tiga lapis.




GAMBAR 1
*Smart Home*

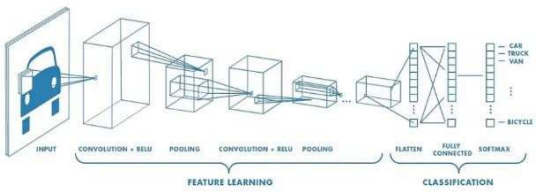
b. Deep Neural Network

DNN mengembangkan representasi berdasarkan data yang dimasukkan dan serangkaian node dan lapisan yang saling terkait dilatih untuk mempelajari fitur yang relevan. DNN sering kali telah dilatih sebelumnya dengan RBM untuk menetapkan titik awal jaringan dalam, namun setelah dibuat, berbagai bentuk DNN dapat menggunakan metode seperti propagasi mundur untuk proses end-to-net yang mulus. mengakhiri pelatihan. Dua bentuk umum DNN adalah CNN dan RNN. Yang pertama adalah yang paling umum karena unggul dalam data spasial/temporal.

$$z^{(l)} = W^{(l)}a^{(l-1)} + b^{(l)} \quad (1)$$

1. Convolutional Neural Network

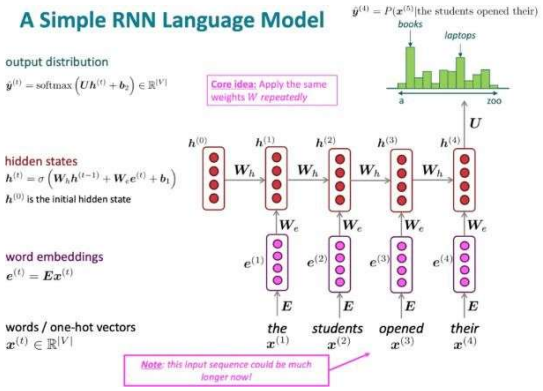
Convolutional Convolutional Neural Network (CNN) adalah salah satu jenis arsitektur jaringan saraf tiruan (Artificial Neural Network, ANN) yang dirancang khusus untuk pengenalan pola dalam data yang memiliki struktur grid, seperti citra atau gambar. CNN memanfaatkan operasi konvolusi untuk mengekstraksi fitur spasial dan hierarkis dari data input. CNN terdiri dari beberapa lapisan utama, yaitu lapisan konvolusi (convolutional layer), lapisan pooling (pooling layer), lapisan aktivasi (activation layer), dan lapisan fully connected (fully connected layer) yang terhubung dengan output. Lapisan konvolusi bertanggung jawab untuk mendeteksi fitur-fitur lokal dari data, sedangkan lapisan pooling bertujuan untuk mengurangi dimensi fitur, sehingga mengurangi kompleksitas komputasi sesuai Gambar.




GAMBAR 2
CNN

2. Reccurant Neural Network

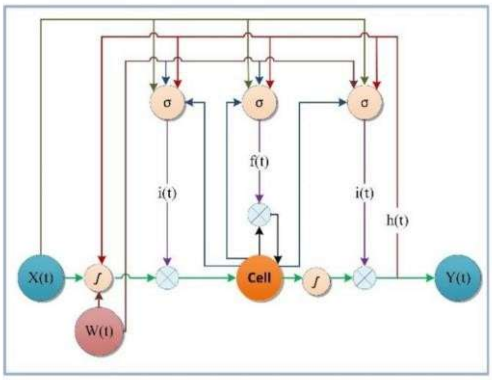
Recurrent Neural Network (RNN) adalah jenis arsitektur jaringan saraf tiruan (Artificial Neural Network, ANN) yang dirancang untuk memproses data berurutan atau data dengan dependensi waktu, seperti deret waktu (time series), teks, atau sinyal suara. Tidak seperti jaringan saraf biasa, RNN memiliki koneksi berulang (recurrent connections), yang memungkinkan informasi dari langkah sebelumnya digunakan untuk memengaruhi pemrosesan langkah berikutnya. Ini memungkinkan RNN untuk mempertahankan "memori" dari data sebelumnya dan memodelkan hubungan temporal dalam data sekuensial.




GAMBAR 2
RNN

3. Long Short-Term Memory

Long Short-Term Memory (LSTM) adalah jenis khusus dari jaringan saraf tiruan yang termasuk dalam keluarga Recurrent Neural Networks (RNN). LSTM dirancang untuk mengatasi masalah vanishing gradient yang sering muncul dalam pelatihan RNN tradisional. Dengan arsitektur khusus yang melibatkan unit memori, LSTM mampu menangkap hubungan jangka panjang dalam data sekuensial seperti teks, audio, atau data waktu.




GAMBAR 4
Desain Blok Sistem

4. Fitur Akustik

Sistem pengenalan ucapan modern umumnya terdiri atas tahap ekstraksi yang bertugas mengubah sinyal suara mentah menjadi representasi yang lebih stabil terhadap perubahan saluran dan kebisingan. Fitur yang paling banyak digunakan berbasis analisis spektral jangka pendek, misalnya Mel-Frequency Cepstral Coefficients (MFCC) dan bank filter mel. Pada pendekatan ini, sinyal suara dipecah menjadi frame pendek, kemudian dihitung spektrum frekuensi menggunakan transformasi Fourier, diproyeksikan ke skala mel yang menyerupai persepsi frekuensi telinga manusia, lalu diambil koefisien cepstral sebagai fitur yang mewakili bentuk spektrum pada tiap frame. Fitur ini terbukti efektif untuk tugas pengenalan ucapan dan menjadi standar de-facto pada berbagai sistem berbasis DNN.

c. Aktuator

Aktuator Relay adalah komponen elektromekanis atau komponen solid-state yang berfungsi sebagai saklar otomatis untuk mengontrol aliran listrik pada perangkat tertentu. Dalam sistem smart home, relay memainkan peran penting dalam menghubungkan dan memutus aliran listrik ke perangkat rumah tangga seperti lampu,

kipas angin, mesin cuci, dan peralatan elektronik lainnya. Relay bekerja dengan menggunakan sinyal listrik kecil dari mikrokontroler atau mikrokomputer (seperti Raspberry Pi atau Arduino) untuk mengendalikan arus listrik yang jauh lebih besar.

d. Mikrokomputer

Mikrokomputer adalah komputer berukuran kecil yang menggunakan mikroprosesor sebagai unit pemrosesan pusat (CPU). Tidak seperti komputer pribadi (PC) yang umumnya terdiri dari banyak komponen yang terpisah, mikrokomputer mengintegrasikan CPU, memori, perangkat input/output (I/O), dan antarmuka lainnya ke dalam satu papan sirkuit tunggal (single-board computer, SBC). Mikrokomputer digunakan secara luas dalam pengendalian perangkat elektronik, pendidikan, penelitian, serta pengembangan perangkat

lunak dan prototipe perangkat keras.

*e. Speaker recognition*

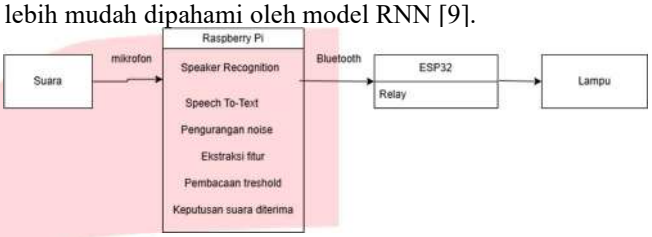
Pada pengenalan pembicara, ciri akustik seperti MFCC sering dikombinasikan dengan model statistik atau jaringan saraf untuk membedakan identitas suara. Beberapa penelitian smart-home menggunakan MFCC yang diekstraksi dari berkas audio oikemudian dilatihkan ke Gaussian Mixture Model (GMM) atau jaringan saraf sebagai model suara tiap penghuni rumah. Sistem akan merekam ucapan pendek dari pengguna baru, menghitung vektor fitur MFCC, dan membangun model pembicara dari fitur tersebut. Saat verifikasi, fitur dari ucapan uji dibandingkan ke seluruh model dan identitas dipilih berdasarkan skor kemiripan tertinggi. Pendekatan ini mampu mencapai akurasi pengenalan pembicara di atas 80% pada skenario keamanan rumah pintar dan menunjukkan bahwa kombinasi fitur spektral dan model pembelajaran mesin cocok digunakan sebagai mekanisme autentikasi tambahan selain pengenalan wajah.

*f. Voice command*

Sistem voice command pada dasarnya adalah sistem yang memproses suara sebagai masukan utama, kemudian menafsirkan makna perintah tersebut dan menghasilkan keluaran yang sesuai, umumnya juga dalam bentuk suara. Arsitektur dasar yang banyak digunakan terdiri dari tiga komponen utama, yaitu speech-to-text (STT) untuk mengubah sinyal ucapan menjadi teks, query processor untuk menganalisis teks dan menentukan aksi atau jawaban yang tepat, serta text-to-speech (TTS) untuk mengubah kembali teks keluaran menjadi sinyal suara yang dapat didengar pengguna. Pendekatan ini memungkinkan pengguna berinteraksi dengan sistem hanya dengan berbicara, tanpa perlu mengetik atau menggunakan antarmuka grafis yang kompleks.

## III. METODE

Dalam rumah pintar, tujuan utama adalah menerapkan sistem pengenalan perintah suara berbasis jaringan saraf di rumah pintar dengan menggunakan kontrol suara untuk mengendalikan perangkat secara otomatis. Sejalan dengan pendekatan oleh Hung dan Giang [9], implementasi ini terdiri dari beberapa tahap utama seperti akuisisi data suara, prapemrosesan sinyal, pelatihan Recurrent Neural Network (RNN), dan integrasi sistem kontrol. Data suara yang digunakan dalam pekerjaan ini terdiri dari beberapa file suara yang berisi perintah dan prompt umum untuk mengendalikan perangkat rumah pintar seperti "nyalakan lampu," "matikan kipas," dan "nyalakan pompa air." Data diambil dalam berbagai kondisi lingkungan untuk menyesuaikan sistem terhadap kebisingan [9]. Setiap rekaman disimpan dalam format WAV pada frekuensi sampling 16 kHz, frekuensi standar dalam pengenalan suara. Data suara dibersihkan dengan penghapusan kebisingan menggunakan algoritma pengurangan spektral, normalisasi amplitudo, dan Ekstraksi Fitur MFCC, memberikan sinyal suara representasi yang lebih mudah dipahami oleh model RNN [9].




GAMBAR 5
Desain Blok Sistem

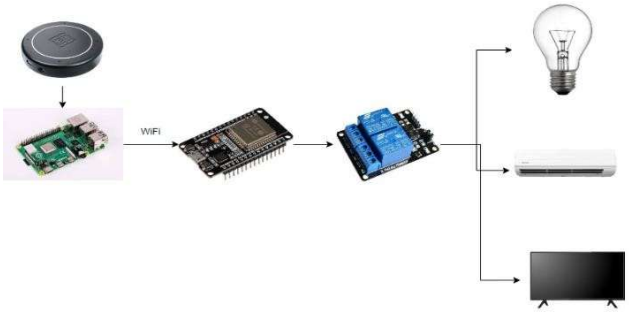
Dalam Gambar 5 terlihat alur cara kerja smart home yang bermula dari input yaitu suara yang ditangkap oleh sensor lalu diteruskan ke mikrokomputer untuk memproses suara yang masuk. Apabila suara yang masuk tidak terdaftar pada profil smart home maka smart home akan menolak menerima perintah, jika suara diterima maka pengguna dapat menyebutkan perintah yang nantinya akan diproses ke mikrtokontroller dan diteruskan ke relay untuk mengontrol pemakaian perangkat listrik.

*Alur Kerja Sistem*

Alur kerja sistem dimulai dari tahap inisialisasi dan pendaftaran suara. Pada tahap ini, seluruh perangkat seperti mikrokomputer, mikrofon, modul komunikasi, mikrokontroler, dan relay terlebih dahulu diaktifkan, kemudian pengguna yang akan diotorisasi diminta merekam beberapa contoh ujaran. Rekaman sinyal audio diproses terlebih dahulu (dibersihkan, dihilangkan derau, dan ekstraksi fitur akustik) dan profil suara (sidik suara) kemudian diekstraksi untuk setiap pengguna, yang disimpan ke dalam basis data sistem. Profil suara ini dapat digunakan sebagai referensi untuk identifikasi dan autentikasi pembicara ketika sistem beroperasi secara normal.

Saat sistem bekerja, sistem berada pada keadaan siaga sambil menunggu perintah berupa suara dari pembicara. Ketika pengguna mengucapkan perintah, mikrofon menangkap sinyal suara dan mengirimkannya ke mikrokomputer untuk diproses. Mikrokomputer melakukan pra-pemrosesan dan ekstraksi fitur, kemudian menjalankan model DNN *speaker recognition* untuk memeriksa apakah suara tersebut berasal dari salah satu pengguna yang sudah terdaftar. Jika hasil kecocokan di bawah nilai ambang yang ditetapkan, perintah akan ditolak dan sistem kembali ke keadaan siaga. Jika suara dikenali dan lulus otentikasi, modul pengenal perintah (*voice command*) menerjemahkan isi ujaran menjadi perintah logika (misalnya “hidupkan lampu ruang tamu” atau “matikan kipas”). Perintah ini dikemas dalam bentuk pesan protokol komunikasi dan dikirimkan dari mikrokomputer ke mikrokontroler (misalnya ESP32). Mikrokontroler kemudian mengaktifkan atau menonaktifkan keluaran digital yang terhubung ke modul relay, sehingga kontak relay berubah posisi dan menghubungkan atau memutuskan beban listrik.

b. *Desain Perangkat Keras*




GAMBAR 6
Desain Perangkat Keras

Desain perangkat keras dalam penelitian ini terlihat dalam Gambar 6 yang memulai sistem dengan mikrofon untuk penangkap suara pengguna yang terhubung ke Raspberry Pi. Setelah di Raspberry Pi memproses suara dan memberi akses pada pembicara maka perkataan perintah yang ditangkap oleh rapberry pi akan di teruskan ke ESP 32 Melalui Bluetooth. Di ESP 32 Setiap kata perintah mempunyai beban yang di program untuk menjalankan input multi-channel relay di dalam rela telah tersambung kabel listrik yang akan menyalurkan daya kepada beban listrik yang akan menyala ataupun mati

c. Fungsi dan fitur

1. Fungsi

Fungsi pengendalian perangkat rumah tangga sistem memungkinkan pengguna untuk mengendalikan berbagai perangkat rumah tangga hanya dengan perintah suara.

Fungsi identifikasi dan otentikasi pembicara sistem tidak hanya mengenali isi perintah, tetapi juga mengidentifikasi siapa yang mengucapkannya melalui modul *speaker recognition* berbasis CNN–BiLSTM.

Fungsi personalisasi di dalam pengaturan karena sistem dapat membedakan tiap pengguna, pengaturan perangkat dapat dipersonalisasi sesuai profil masing-masing penghuni rumah.

Fungsi peningkatan keamanan smart home dengan serintah suara yang dikombinasikan dengan pengenalan suara telah memberikan tingkat keamanan baru untuk perangkat-perangkat ini, misalnya memungkinkan kita mengakses pintu otomatis, kunci elektronik, atau perangkat sensitif.

Fungsi pemantauan dan pencatatan perintah
sistem ini berfungsi untuk mendokumentasikan atau merekam riwayat perintah yang berhasil dijalankan atau gagal, misalnya, karena pembicara tidak dikenali.

2. Fitur

Multi-user voice profile Sistem mampu menyimpan dan mengelola profil suara dari beberapa pengguna sekaligus.

Identifikasi dan otentikasi berbasis suara Pada saat menerima perintah, sistem melakukan identifikasi dan sekaligus otentikasi pembicara.

Penanganan kebisingan (noise cancellation dan filtering) Sistem ini dilatih dengan pra-pemrosesan seperti sanitasi gelombang, pengurangan kebisingan, dan pemilihan fitur akustik untuk kebisingan non-ucapan guna meningkatkan keandalan pengenalan di lingkungan rumah yang bising.

Pemrosesan lokal dengan model DNN Seluruh proses pengenalan pembicara dilakukan secara lokal pada Raspberry Pi menggunakan model DNN (CNN– BiLSTM dengan mekanisme attention) tanpa ketergantungan pada layanan cloud.

Respon waktu nyata (real-time) Desain sistem menangani sinyal suara, ekstraksi fitur, dan inferensi model dalam waktu yang relatif singkat.

Mekanisme penolakan perintah (reject unknown speaker) Sistem ini dilengkapi dengan mekanisme penolakan perintah, yaitu perintah akan ditolak jika pembicara tidak dikenali, atau jika tingkat kesamaan suara berada di bawah ambang batas yang ditentukan.

d. Flowchart sistem

**Gambar 7** Flowchart Sistem

Alur kerja keseluruhan sistem ditunjukkan pada flowchart Gambar 7. Proses dimulai dari tahap inisiasi sistem, yaitu ketika mikrokomputer menyalakan seluruh perangkat, memuat model speaker recognition, serta membangun

koneksi dengan modul ESP32 dan rangkaian relay. Setelah inisiasi, sistem memiliki dua mode utama, yaitu pendaftaran pembicara (enrolment) dan operasi verifikasi + pengendalian smart home. Pada mode pendaftaran, pengguna baru memilih kata sandi suara tertentu, kemudian diminta mengucapkan frasa tersebut berulang sebanyak lima kali. Setiap rekaman diproses melalui pra-pemrosesan dan ekstraksi fitur akustik, lalu digabungkan untuk membentuk profil suara (enrol) pengguna. Jika kualitas rekaman dinilai memadai, profil tersebut disimpan sebagai model pembicara terdaftar yang nantinya digunakan pada proses verifikasi.

Pada mode operasi, pengguna terlebih dahulu memberikan perintah verifikasi pembicara. Sistem menangkap suara, mengekstraksi fitur, dan membandingkannya dengan profil yang tersimpan. Tahap "penyesuaian verifikasi suara" pada flowchart merepresentasikan proses pengecekan apakah skor kemiripan suara melewati nilai ambang yang telah ditentukan; jika belum memenuhi, pengguna diminta mengulangi verifikasi, sedangkan jika berhasil pengguna dinyatakan terverifikasi. Setelah verifikasi berhasil, pengguna dapat mengucapkan perintah smart home (misalnya menyalakan atau mematikan lampu). Ujaran tersebut diubah menjadi teks oleh modul speech-to-text dan dipetakan ke perintah logika sistem. Perintah kemudian diteruskan ke modul ESP32, yang selanjutnya mengendalikan modul relay sesuai beban listrik yang terhubung sehingga perangkat rumah tangga menyala atau mati sesuai perintah. Sistem terus berada pada keadaan aktif selama masih ada perintah; apabila tidak ada perintah yang diterima dalam jangka waktu sekitar lima menit, sesi dikembalikan ke keadaan siaga dan alur kembali ke awal.

## IV. HASIL DAN PEMBAHASAN

### *A.* Realisasi Alat

Perangkat lunak dan perangkat keras digunakan untuk mendefinisikan sistem. Sistem ini bergantung pada sensor suara dari mikrofon dalam antarmuka rumah pintar dengan Raspberry Pi 5 untuk komputasi dan kinerja pemrosesan DNN. Untuk merekam suara pengguna ke dalam mikrofon, Raspberry Pi yang terhubung ke mikrofon digabungkan untuk mendapatkan data dari mikrofon dan digunakan dalam proses pembelajaran mendalam untuk menganalisis suara dan mengidentifikasi pembicara. Sistem dioperasikan dengan perangkat lunak dan perangkat keras. Dalam sistem ini, di mana seseorang menggunakan mikrofon sebagai sensor suara di rumah pintarnya yang terhubung ke Raspberry Pi 5 untuk komputasi beberapa proses DNN. Raspberry Pi berkomunikasi dengan mikrofon untuk mengumpulkan suara dari pengguna dan memprosesnya dalam model pembelajaran mendalam untuk mendeteksi pembicara dan menentukan identitas mereka. Saat pengaktifan terlebih dahulu megucapkan kata kunci untuk mendapat akses ke smart home. Sistem ini dapat menambah suara yang dapat mengakses smart home dengan melakukan penyebutan kata kunci dan setelah itu akan mengambil 5 sampel suara untuk pengenalan pembicara dengan mengucapkan kalimat selama 1-2 detik. Setelah dapat mengetahui pembicara maka proses akan dilanjutkan ke ESP32 melalui komunikasi nirkabel bluetooth untuk proses pengaktifan smart home, ESP32 berperan sebagai aktuator dalam sistem ini dengan menangkap sinyal dari Raspberry pi yang memberikan perintah berupa speech to text lalu diteruskan ke ESP32 yang didalamnya terdapat proses untuk memberi perintah ke relay quad-channel yang dapat mengaktifkan dan mematikan perangkat rumah tangga. Di Gambar 8 merupakan foto rangkaian sistem.


GAMBAR 8
Alat Smart Home

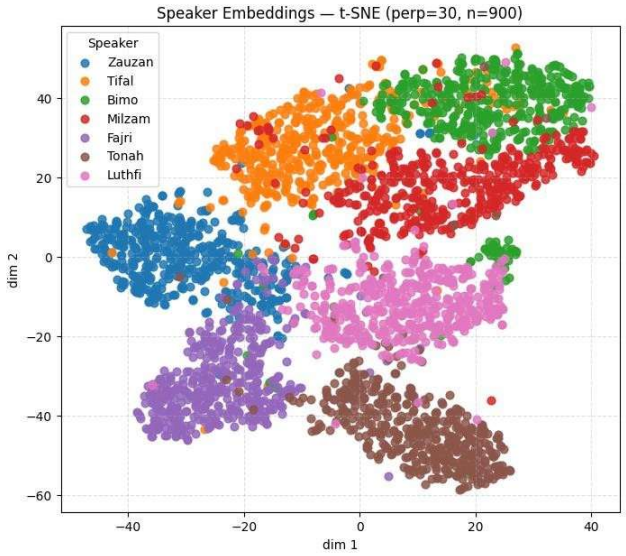
### B. Training DNN

Algortima DNN yang dipakai dalam penelitian ini merupakan CNNSpeaker dan RNN BiLSTM. Kedua algoritma tersebut digabung menjadi satu model yang dikembangkan untuk membaca setiap profil suara per pembicara, seriap pembicara mempunyai suara yang berbeda-beda. . Pada bagian awal, lapisan CNN berperan mengekstraksi pola-pola lokal pada domain waktu–frekuensi dari fitur akustik 105 dimensi (mel-spectrogram, MFCC, delta, delta-delta, RMS, dan f_0), sehingga karakteristik spektral yang khas dari suara masing-masing orang dapat ditangkap secara lebih terstruktur. Hasil ekstraksi CNN kemudian diteruskan ke lapisan BiLSTM yang mampu memodelkan urutan frame fitur dari dua arah (forward dan backward), sehingga informasi temporal seperti dinamika intonasi, ritme bicara, dan perubahan energi suara turut dipertimbangkan.

| *Label* | *Precision (%)* | *Recall (%)* | *F1-score (%)* | *Support (%)* |
|---|---|---|---|---|
| Thifal | 88,46 | 88,46 | 88,40 | 14,2 |
| Zauzan | 90,00 | 91,41 | 90,70 | 14,2 |
| Bimo | 84,78 | 90,00 | 87,31 | 14,2 |
| Milzam | 91,60 | 85,16 | 88,26 | 14,2 |
| Fajri | 94,26 | 90,55 | 92,37 | 14,2 |
| Tonah | 97,58 | 93,80 | 95,65 | 14,2 |
| Luthfi | 90,30 | 94,53 | 92,37 | 14,2 |
| Accuracy | 91 | | | |

TABEL 2
*Confussion matrix*

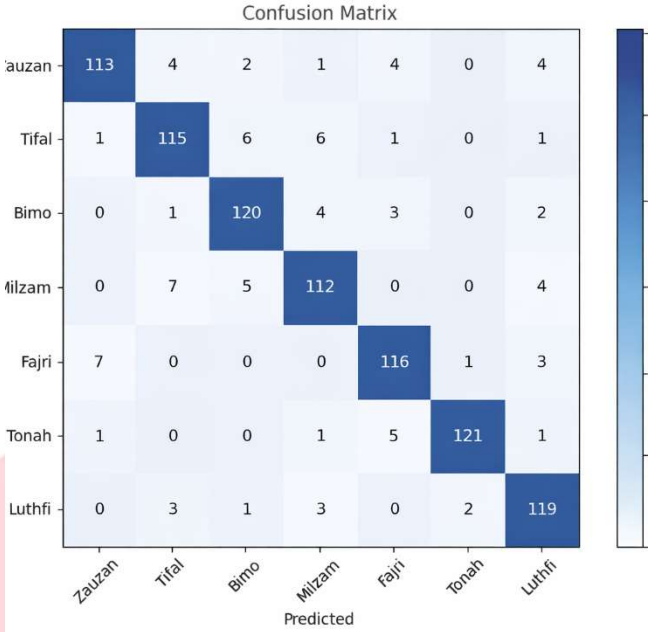
Confusion matrix pada Tabel 1 memperlihatkan sebaran prediksi model pengenalan pembicara untuk tujuh kelas, yaitu Bimo, Fajri, Luthfi, Tonah, Milzam, Tifal, dan Zauzan. Nilai diagonal pada matriks menunjukkan jumlah prediksi benar untuk masing-masing pembicara, sedangkan nilai di luar diagonal menunjukkan kesalahan klasifikasi (suara pembicara A diprediksi sebagai pembicara B). Secara keseluruhan, sistem menghasilkan akurasi 91% pada 900 sampel uji, dengan nilai macro dan weighted average precision, recall, dan F1-score yang semuanya berada di sekitar 0,91, yang menandakan performa model relatif seimbang antar kelas. MAE suhu rata-rata < 0.5°C, sehingga model cukup presisi untuk memprediksi dinamika termal ruangan.




GAMBAR 9
*Embedding* Pembicara

Gambar 9 menunjukkan visualisasi embedding pembicara menggunakan t-SNE (perplexity = 30, n = 900). Masing-masing titik merepresentasikan satu contoh ujaran, sedangkan perbedaan warna menyatakan identitas pembicara. Terlihat bahwa beberapa titik dengan warna yang sama cenderung berkumpul secara lokal, namun antar cluster pembicara masih saling bercampur sehingga pemisahan ruang fitur belum sepenuhnya tegas. Hal ini menunjukkan bahwa model DNN telah mampu menangkap karakteristik suara masing-masing pembicara sampai tingkat tertentu, tetapi masih terdapat kemiripan embedding pada beberapa pembicara yang berkontribusi terhadap kesalahan klasifikasi yang tampak pada confusion matrix.

Confusion matrix pada Gambar 10 menggambarkan kemampuan sistem dalam membaca (mengenali) setiap pembicara berdasarkan suara yang diujikan. Sumbu vertikal menunjukkan pembicara sebenarnya (actual), sedangkan sumbu horizontal menunjukkan pembicara hasil prediksi model (predicted). Angka pada diagonal utama (kiri atas ke kanan bawah) merupakan jumlah True Positive (TP), yaitu banyaknya sampel suara yang dikenali dengan benar sebagai pemiliknya. Angka di luar diagonal adalah kesalahan klasifikasi. False Negative (FN) untuk suatu pembicara = sampel yang sebenarnya milik pembicara tersebut, tetapi diprediksi sebagai pembicara lain (menyebar di baris yang sama). False Positive (FP) = sampel pembicara lain yang justru diprediksi sebagai pembicara tersebut (menyebar di kolom yang sama). Semakin besar nilai diagonal dan semakin




GAMBAR 10
*Confussion matrix*

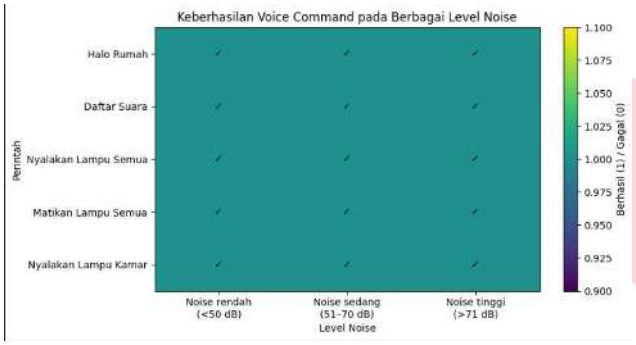
Pengujian
Pengujian Terhadap Pembicara

| Pendaftar | Pembicara | Kata yang terdeteksi | Treshold | Dikenali |
|---|---|---|---|---|
| Terdaftar | Bimo | Halo Rumah | 0.7859 | Ya |
| Terdaftar | Bimo | Halo | 0.8521 | Ya |
| Terdaftar | Bimo | Halo Rumah | 0.8412 | Ya |
| Terdaftar | Bimo | Kalau Rumah | 0.9101 | Ya |
| Terdaftar | Bimo | Halo Rumah | 0.8267 | Ya |
| Tidak Terdaftar | Luthfi | Halo Rumah | 0.515 | Tidak |
| Tidak Terdaftar | Luthfi | Rumah | 0.6573 | Tidak |
| Tidak Terdaftar | Luthfi | Halo | 0.2632 | Tidak |
| Tidak Terdaftar | Luthfi | Halo Rumah | 0.6721 | Tidak |
| Tidak Terdaftar | Luthfi | Halo Rumah | 0.7193 | Ya |

TABEL 2
*Pengujian Speaker Recognition*

Tabel pengujian speaker recognition pada Tabel 2 memperlihatkan respon sistem terhadap satu pembicara yang terdaftar (Bimo) dan satu pembicara tidak terdaftar (Luthfi) dengan batas nilai ambang (threshold) skor kecocokan sebesar 0,7. Untuk pembicara terdaftar, kelima percobaan yang dilakukan dengan variasi pengucapan kata kunci ("Halo Rumah", "Halo", dan "Kalau Rumah") menghasilkan skor kecocokan antara 0,7859–0,9101, semuanya berada di atas threshold dan dikategorikan "Ya" (dikenali). Hal ini menunjukkan bahwa sistem mampu menerima pembicara sah secara konsisten, tanpa terjadi false rejection pada skenario uji ini, dengan rata-rata skor kecocokan sekitar 0,84 yang relatif jauh dari batas 0,7.

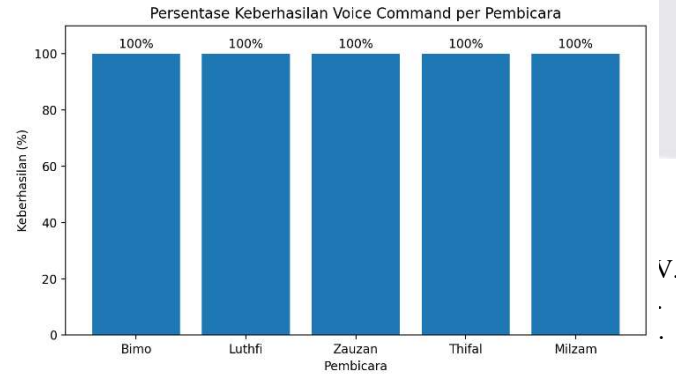
2. Pengujian Perhadap Noise




GAMBAR 11
Pengujian Terhadap Noise

Gambar pengujian terhadap noise pada memperlihatkan keberhasilan sistem dalam mengeksekusi beberapa perintah kunci ("Halo Rumah", "Daftar Suara", "Nyalakan Lampu Semua", "Matikan Lampu Semua", dan "Nyalakan Lampu Kamar") pada tiga kondisi kebisingan lingkungan, yaitu noise rendah (< 50 dB), noise sedang (51–70 dB), dan noise tinggi (> 71 dB). Tanda centang (✓) yang muncul pada setiap kombinasi perintah dan level noise menunjukkan bahwa seluruh skenario pengujian tersebut dapat dikenali dan dieksekusi dengan benar oleh sistem: kata pemicu ("Halo Rumah") tetap terdeteksi, daftar suara dapat ditampilkan, dan perintah pengendalian lampu (menyalakan maupun mematikan, baik per ruangan maupun keseluruhan) berhasil mengubah status beban listrik sesuai harapan pada ketiga tingkat kebisingan..

*3.* Pengujian *Voice command*




GAMBAR 12
*Voice Command*

Dari Gambar 12 terlihat bahwa sebagian besar kombinasi pembicara dan perintah menghasilkan tanda centang, yang berarti hampir semua perintah yang diujikan kepada lima pembicara otorisasi dapat dijalankan oleh sistem secara konsisten. Perintah dasar seperti menghidupkan dan mematikan lampu di tiap ruangan, serta perintah agregat "Hidupkan Semua" dan "Matikan Semua", dapat direspons dengan benar oleh lebih dari satu pembicara, sehingga menunjukkan bahwa modul speech-to-text, pemetaan perintah, serta pengendalian aktuator berjalan sesuai rancangan. Beberapa sel yang kosong mengindikasikan bahwa kombinasi tersebut belum diuji atau tidak berhasil dieksekusi pada saat pengujian, namun secara umum tingkat keberhasilan fungsional voice command pada prototipe ini mendekati penuh. Hasil ini menegaskan bahwa integrasi antara modul pengenalan pembicara, pengenalan perintah suara, dan sistem kendali smart home sudah bekerja dengan baik pada kondisi uji coba di lingkungan rumah..

D. Analisis Keseluruhan

Secara keseluruhan, sistem smart home berbasis perintah suara dengan speaker recognition yang dikembangkan telah memenuhi tujuan penelitian, yaitu mengenali pembicara terdaftar, menolak pembicara tidak terdaftar, serta menjalankan perintah pengendalian beban listrik secara stabil pada berbagai kondisi uji. Kinerja model DNN CNN–BiLSTM menunjukkan performa yang baik dengan akurasi sekitar 91% pada 900 sampel uji dan nilai precision, recall, serta F1-score rata-rata sebesar 0,91, yang menandakan kemampuan identifikasi pembicara cukup konsisten antar kelas. Kesalahan klasifikasi yang masih muncul umumnya terjadi pada pembicara dengan karakteristik suara yang mirip, sehingga beberapa pasangan pembicara tertentu lebih sering tertukar, namun tidak sampai menurunkan performa sistem secara signifikan.

Dari sisi keamanan, penggunaan threshold 0,7 terbukti mampu menjadi mekanisme verifikasi yang efektif. Pada pengujian multi-pembicara, sistem dapat menerima pengguna terdaftar dan menolak pengguna tidak terdaftar secara konsisten (FAR/FRR rendah), meskipun pada skenario tertentu masih terdapat peluang *false acceptance* yang menunjukkan perlunya evaluasi threshold jika sistem digunakan pada kondisi nyata yang lebih bervariasi. Pada aspek fungsional, modul voice command berhasil mengeksekusi seluruh perintah utama (kendali lampu per ruangan dan perintah global) serta tetap berjalan pada variasi tingkat kebisingan, yang menunjukkan bahwa pra- pemrosesan (sanitize dan denoise) membantu menjaga kualitas sinyal masukan. Namun demikian, hasil pengujian waktu respons menunjukkan sistem masih memiliki latensi total sekitar 10–14 detik dari kata kunci hingga aktuator bekerja, sehingga sistem dinilai sudah layak untuk prototipe, tetapi masih dapat ditingkatkan melalui optimasi proses STT, inferensi model, dan komunikasi Raspberry Pi–ESP32 agar respons lebih cepat dan terasa lebih real-time.

V. KESIMPULAN DAN SARAN

. Kesimpulan

. Implementasi arsitektur CNN–BiLSTM dengan fitur akustik 105 dimensi (log Mel-spectrogram, MFCC, delta, delta-delta, RMS, dan $f_0$) menghasilkan akurasi klasifikasi sekitar **91%** pada 900 sampel uji dengan

nilai precision, recall, dan F1-score rata-rata (macro dan weighted) juga sebesar 0,91. Hal ini menunjukkan bahwa tujuan utama penerapan DNN sebagai pemetaan nonlinier dari fitur akustik ke identitas pembicara telah tercapai dengan baik.

2. Analisis confusion matrix dan grafik per kelas menunjukkan bahwa seluruh pembicara memiliki F1- score di atas 87%, dengan kelas terbaik mencapai sekitar 96%. Kesalahan klasifikasi yang muncul umumnya terkonsentrasi pada beberapa pasangan pembicara dengan karakteristik suara yang mirip, sementara tidak ada kelas yang performanya jatuh drastis. Ini menandakan bahwa sistem mampu membaca dan memisahkan profil suara masing-masing pembicara secara cukup konsisten.
3. Pengujian verifikasi dengan threshold skor kecocokan sekitar **0,7** menunjukkan bahwa sistem dapat menerima seluruh percobaan dari pembicara terdaftar (FRR ≈ 0%) dan, pada skenario empat pembicara (dua terdaftar dan dua tidak terdaftar), menolak seluruh percobaan dari pembicara yang tidak terdaftar (FAR ≈ 0%). Pada skenario yang lebih ketat masih ditemukan satu kasus false acceptance, tetapi secara umum threshold ini sudah memberikan kompromi yang baik antara keamanan dan kenyamanan.

yang lebih ringan atau dijalankan lokal penuh), meminimalkan overhead I/O dan serialisasi data, serta

b. Saran

Untuk meningkatkan generalisasi model dan mengurangi kesalahan pada pasangan pembicara yang masih sering tertukar, disarankan menambah jumlah data latih per pembicara, menambah variasi intonasi,

kecepatan bicara, dan kondisi rekaman, serta memperluas jumlah pembicara (misalnya > 10 orang) sehingga model lebih robust terhadap perbedaan karakter suara di dunia nyata.

1. Di masa mendatang dapat dieksplorasi arsitektur yang lebih spesifik untuk speaker recognition, seperti x- vector, ECAPA-TDNN, atau kombinasi CNN– Transformer, serta penggunaan loss function metric-

learning (ArcFace, AAM-Softmax) untuk memperbesar jarak antar embedding pembicara. Hal ini berpotensi untuk meningkatkan pemisahan cluster di ruang fitur dan menurunkan False Acceptance/False Rejection.

2. Waktu respons dapat dipersingkat dengan mengoptimasi modul speech-to-text (misalnya model mengoptimalkan jalur komunikasi Raspberry Pi– ESP32 (misalnya protokol yang lebih ringan atau pengiriman perintah yang lebih ringkas). Penggunaan akselerasi hardware (GPU kecil / NPU jika tersedia) juga bisa dipertimbangkan untuk mempercepat inferensi DNN.

## REFERENSI

[1]M. R. Alam, M. B. I. Reaz, dan M. A. M. Ali, “A review of smart homes - Past, present, and future,” IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, vol. 42, no. 6, 2012, doi: 10.1109/TSMCC.2012.2189204.

[2]N. S. Nugroho dan B. A. Pramudita, “Sistem Smart Home dengan Voice Command Untuk Memantau dan Mengaktifkan Peralatan Listrik Rumah.”

[3]A. Iqbal dkk., “Interoperable Internet-of-Things platform for smart home system using Web-of-Objects and cloud,” Sustain Cities Soc, vol. 38, 2018, doi: 10.1016/j.scs.2018.01.044.

[4]T. K. Tran, K. T. Huynh, D. N. Le, M. Arif, dan H. M. Dinh, “A Deep Trash Classification Model on Raspberry Pi 4,” Intelligent Automation and Soft Computing, vol. 35, no. 2, hlm. 2479–2491, 2023, doi: 10.32604/iasc.2023.029078.

[5]S. Yin dkk., “Noisy training for deep neural networks in speech recognition,” EURASIP J Audio Speech Music Process, vol. 2015, no. 1, 2015, doi: 10.1186/s13636-014-0047-0.

[6]N. Saxena dan D. Varshney, “Smart Home Security Solutions using Facial Authentication and Speaker Recognition through Artificial Neural Networks,” International Journal of Cognitive Computing in Engineering, vol. 2, 2021, doi: 10.1016/j.ijcce.2021.10.001.

[7]F. M. Talaat, “Explainable Enhanced Recurrent Neural Network for lie detection using voice stress analysis,” Multimed Tools Appl, vol. 83, no. 11, 2024, doi: 10.1007/s11042-023-16769-w.

[8]S. Ilham, A. Muin, dan D. Candro, “System control device electronics smart home using neural networks,” Advances in Science, Technology and Engineering Systems, vol. 2, no. 5, 2017, doi: 10.25046/aj020507.

[9]P. D. Hung, T. M. Giang, L. H. Nam, dan P. M. Duong, “Vietnamese speech command recognition using Recurrent Neural Networks,” International Journal of Advanced Computer Science and Applications, vol. 10, no. 7, 2019, doi: 10.14569/ijacsa.2019.0100728..