Jurnal Teknologi Informasi dan Pendidikan
Volume 18, No. 2, September 2025
https://doi.org/10.24036/jtip.v18i2.971

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

A Web-Based SIBI Sign Language Translator Application with Speech-to-Text Feature Using CNN and MediaPipe

Fathur Rahman[1*]✉, Ahmaddul Hadi[1], Dony Novaliendry[1], Agariadne Dwinggo Samala[1]
[1]Faculty of Engineering, Universitas Negeri Padang, Padang, Indonesia
Corresponding Author: fathur@student.unp.ac.id

Article Information

Article history:

No. 971
Rec. June 03, 2025
Rev. September 07, 2025
Acc. September 07, 2025
Pub. September 09, 2025
Page. 1033 – 1042

Keywords:

- *Sign language translator*
- *SIBI*
- *Hand gesture recognition*
- *Convolutional neural network*
- *Speech-to-text*
- *MediaPipe*
- *Web application*

ABSTRACT

This study developed a web-based application to facilitate two-way communication between individuals with hearing impairments and the general public. The application translated hand gestures based on the Indonesian Sign System into text using a Convolutional Neural Network model and real-time landmark detection. Additionally, it converted spoken language into text through speech recognition technology, which was then displayed alongside the corresponding sign language images.

The system used a camera to capture hand gestures, which were processed into landmark data and classified into letters A to Z. Voice input was processed directly in the browser without additional installations. The application was designed to be lightweight, interactive, and compatible with various devices.

Testing results showed that the gesture recognition feature achieved high accuracy, ranging from 98.71% to 100%. The speech-to-text feature also provided accurate transcription results, both for individual letters and complete sentences. Accuracy decreased at distances beyond 30 cm and in noisy environments.

The integration of gesture recognition and speech-to-text conversion in a single web platform offered an effective, accessible, and inclusive communication solution for users with special needs.

How to Cite:

Rahman, F., & et al. (2025). A Web-Based SIBI Sign Language Translator Application with Speech-to-Text Feature Using CNN and MediaPipe. Jurnal Teknologi Informasi Dan Pendidikan, 18(2), 1033-1042. https://doi.org/10.24036/jtip.v18i2.971

1. INTRODUCTION

Communication is a vital element in human life that allows individuals to exchange information and build social relationships[1]. However, individuals with hearing

impairments often face significant communication barriers, especially when interacting with people who do not understand sign language. This creates a communication gap that affects their accessibility, social participation, and overall quality of life.

In Indonesia, the sign language used by the deaf community is Sistem Isyarat Bahasa Indonesia (SIBI). SIBI is systematically designed based on Indonesian grammar and utilizes hand gestures, facial expressions, and body posture as mediums for conveying messages [2]. Despite being standardized, public understanding of SIBI remains low, highlighting the need for technological innovations to bridge the communication gap between sign language users and non-users.

The advancement of artificial intelligence technology presents new opportunities for creating inclusive communication systems. Convolutional Neural Network (CNN) is one of the deep learning methods that is effective in recognizing visual patterns such as hand gestures. Several studies have proven its effectiveness in mobile-based [3] and virtual-reality-based sign language applications [4]. To support accurate and real-time detection, this system also utilizes MediaPipe, a library developed by Google capable of extracting hand coordinates in the form of three-dimensional landmarks [5].

The CNN architecture consists of several layers such as convolutional, pooling, and fully connected layers that play an essential role in effectively recognizing spatial patterns [6]. The application of deep learning in computer vision systems has led to significant advances in pattern recognition supported by large datasets and GPU-based training [7].

In addition to the visual aspect, voice recognition technology also plays an important role in facilitating two-way communication. The Web Speech API enables voice-to-text conversion directly in the browser without requiring additional devices [8]. This approach is supported by studies that highlight the relevance of speech recognition technology for hearing-impaired users, especially under varying acoustic conditions [9]. With this technology, the voice of non-deaf users can be converted into text and then displayed as corresponding SIBI sign images.

This research aims to develop a web application that can translate SIBI sign language into text and vice versa, by converting voice into text and displaying appropriate SIBI sign images. By integrating CNN, MediaPipe, and the Web Speech API, this application is expected to become an innovative solution that enhances communication between individuals with hearing impairments and the general public in a more effective and inclusive manner.

2. RESEARCH METHOD

This study designs a web application for translating SIBI sign language with Speech-to-Text functionality and hand gesture detection based on CNN and MediaPipe. The system is designed to enable two-way communication, by recognizing voice input using the Web Speech API and detecting hand movements through MediaPipe, which are then classified

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

by a CNN model [10] [8][12]. The Waterfall method was chosen as it is suitable for system development with a linear workflow and clearly defined requirements from the outset [13]

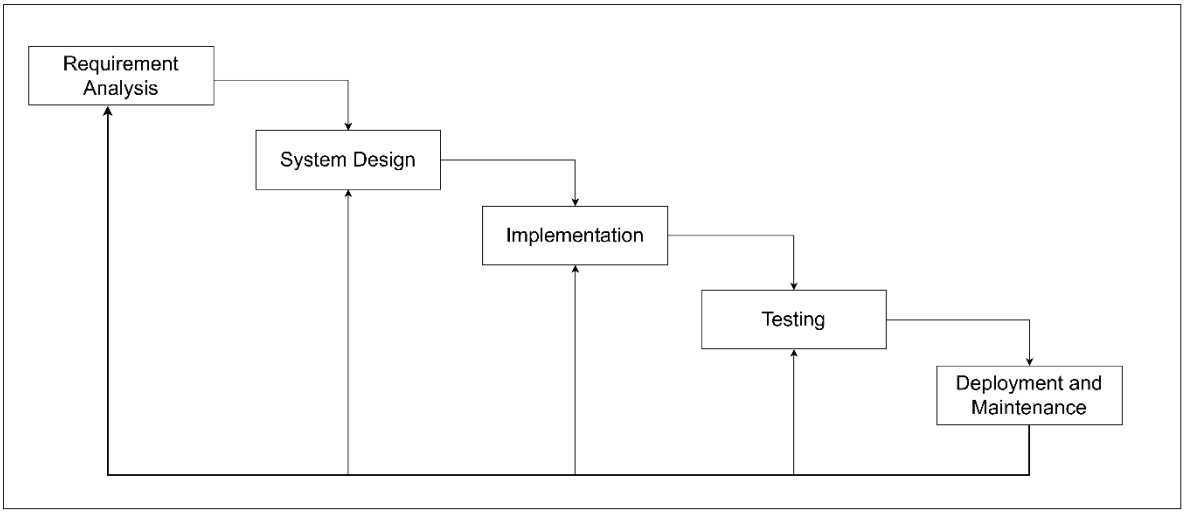

Figure 1. Waterfall Model in System Development

2.1. Requirement Analysis

Needs analysis aims to identify the core functions of the system to enhance communication between individuals with hearing impairments and the general public. The system will detect hand movements in real-time via a camera and classify gestures into letters A–Z using a CNN model, with input in the form of hand landmark coordinates provided by the MediaPipe library [11][14].

In addition, the system converts user speech into text directly using the Web Speech API on the client side, enabling efficient speech-to-text processing without the need for additional installations [8]. Both gesture recognition results and speech-to-text outputs are displayed visually—in the form of text and corresponding SIBI sign language images. This web-based application is designed to be lightweight, interactive, and responsive so that it can be accessed across various devices without requiring special installations [15].

2.2. System Design

System design includes the architecture and workflow design for each feature. The system architecture is divided into two main modes:

Jurnal Teknologi Informasi dan Pendidikan
Volume 18, No. 2, September 2025
https://doi.org/10.24036/jtip.v18i2.971

2.2.1. Mode 1 Speech to Text

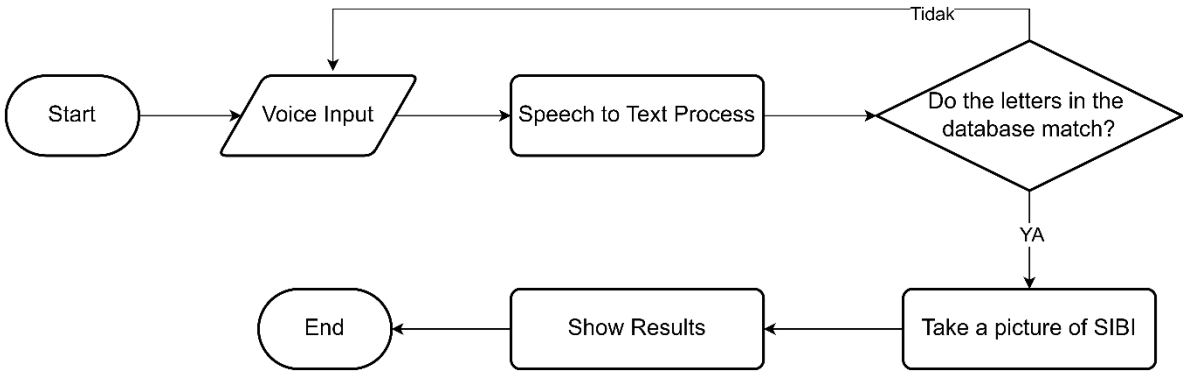

Figure 2. Flowchart Speech to Text

Figure 2 illustrates the system workflow, which begins with activating the camera to capture the user's hand gestures based on SIBI (Indonesian Sign System). The system then extracts hand landmark points using MediaPipe. The extracted landmark data are subsequently processed and reshaped to match the input format required by the Convolutional Neural Network (CNN) model. If the detection confidence score is sufficiently high, the system proceeds with prediction using the CNN model and displays the translated output. Conversely, if the confidence score is low, the prediction process is not carried out. The process concludes after the result is displayed to the user.

2.2.2. Mode 2 SIBI Sign Language Detection

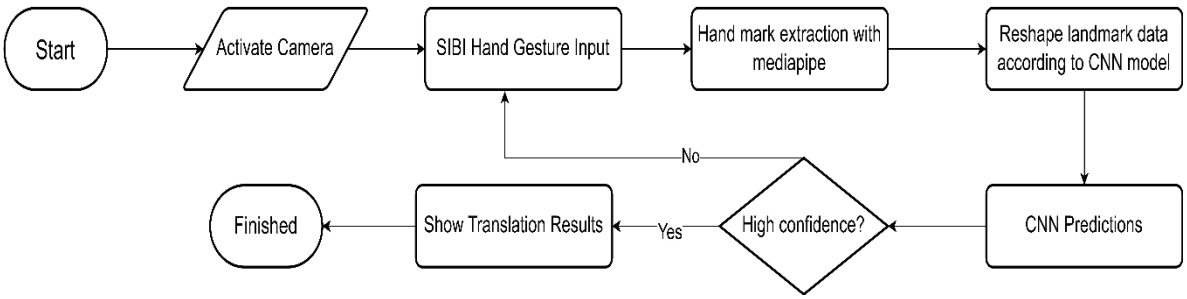

Figure 3. Flowchart SIBI Sign Language Detection

Figure 2. Flowchart of the hand gesture recognition system for SIBI translation. The process begins with the activation of the camera to capture the user's hand movements. The system then extracts hand landmark points using MediaPipe. The obtained landmark data are reshaped to match the input format required by the CNN model. If the confidence score from the detection is sufficiently high, the system proceeds with prediction using CNN and displays the translated result. If the confidence score is low, the prediction is not executed. The process ends once the result is displayed.

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

2.3. Implementation

The system is developed using various technological components. Flask is used as the backend framework due to its compatibility with Python-based machine learning libraries [15]. Hand gesture coordinate extraction is performed using MediaPipe, a library capable of detecting hand landmarks in real-time [11]. For the classification of alphabets A–Z, a Convolutional Neural Network (CNN) model is employed, which is effective in recognizing spatial patterns [14]. Speech-to-text conversion is handled by the Web Speech API, a JavaScript-based tool that enables processing directly in the browser [8]. A MySQL database stores SIBI letter images, which are displayed as the output of the detection process ra[10], 2023). The user interface is designed using HTML, CSS, and JavaScript.

The dataset consists of 52,000 hand gesture samples (2,000 samples for each letter A–Z), used to train the CNN model. The data is stored in CSV format, based on the hand landmark extraction results from MediaPipe (Koyineni et al., 2024).

2.4. Integration and Testing

System testing is conducted to ensure that the main functions operate according to the specifications. The Black Box Testing method is used for:

1) Gesture Recognition, which tests the accuracy of the CNN model in recognizing SIBI letters (A–Z) based on hand coordinates from MediaPipe.
2) Speech-to-Text, which tests the system's success in converting speech to text using the Web Speech API and matching it with the displayed sign language images.

Additionally, supplementary testing is carried out under special conditions to evaluate system robustness, including:

1) Variation in hand distance from the camera for the sign language detection feature.
2) Noisy environment conditions for the Speech-to-Text feature.

2.5. Deployment

After the system was successfully implemented and tested, the application was deployed to a server to enable online access without requiring installation. Users can directly utilize the translator features through a browser, both on desktop and mobile devices.

3. RESULTS AND DISCUSSION

System testing is conducted to ensure that each main function of the system operates according to specifications. The Black Box Testing method is used to test two main features: hand gesture recognition and speech-to-text conversion. Additionally, supplementary tests are carried out under special conditions to evaluate system robustness, such as variations in

hand distance to the camera and testing in noisy environments for the Speech-to-Text feature.

3.1. Gesture Recognition Feature Testing

The gesture recognition feature is designed to recognize hand movements representing SIBI letters from A to Z. The system uses MediaPipe to capture hand landmark coordinates and a CNN to classify them into letters

Testing is conducted by providing hand gesture inputs for each letter one by one using a camera. The system is tested from a functional perspective, specifically whether it can process the gesture input and display the detected letter on the screen.

Table 1. Blackbox testing results

No	Letter	Accuracy (%)	Status	No	Letter	Accuracy (%)	Status
1	A	100.00	**Succes**	14	N	100.00	**Succes**
2	B	100.00	**Succes**	15	O	100.00	**Succes**
3	C	100.00	**Succes**	16	P	99.99	**Succes**
4	D	100.00	**Succes**	17	Q	100.00	**Succes**
5	E	99.77	**Succes**	18	R	99.94	**Succes**
6	F	99.91	**Succes**	19	S	99.61	**Succes**
7	G	100.00	**Succes**	20	T	100.00	**Succes**
8	H	100.00	**Succes**	21	U	99.99	**Succes**
9	I	100.00	**Succes**	22	V	100.00	**Succes**
10	J	98.71	**Succes**	23	W	100.00	**Succes**
11	K	100.00	**Succes**	24	X	100.00	**Succes**
12	L	100.00	**Succes**	25	Y	100.00	**Succes**
13	M	100.00	**Succes**	26	Z	99.34	**Succes**

Based on Table 1, the gesture recognition feature demonstrates high accuracy for all SIBI letters, with accuracy values ranging from 98.71% to 100%. These results indicate that the system is capable of recognizing gesture patterns well, although there is a slight decrease in accuracy for certain letters such as “J” and “Z.”

3.2. Speech-to-Text Feature Testing

This feature converts voice input into text in real-time using the Web Speech API, then matches it with a database of SIBI images.

Testing is conducted by randomly pronouncing letters A–Z through a microphone. The system is evaluated on its ability to:

1) Accurately convert speech to text.
2) Display the corresponding SIBI letter image based on the recognized text.

Table 2. Blackbox testing results Speech to text

No	Speech Input	Transcribed Text	SIBI Image Displayed	Status
1	A	A	Yes	Success
2	D	D	Yes	Success

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

3	P	P	Yes	Success *
4	Z	Z	Yes	Success
5	Please open the door	Please open the door	Yes	Success
6	I am hungry	I am hungry	Yes	Success
7	Good morning	Good morning	Yes	Success
8	I want to eat now	I want to eat now	Yes	Success
9	I learn sign language at school	I learn sign language at school	Yes	Success
10	How are you, I hope you're always well	How are you, I hope you're always well	Yes	**Succes**

Based on Table 2, the system successfully converts speech into text and displays the corresponding SIBI letter images. All tests show accurate transcription results, both for single letters and full sentences.

3.3. Testing Based on Hand to Camera Distance

Table 3. Testing Based on Hand Distance to Camera.

No	Letter	15 cm	30 cm	50 cm
1	A	100%	Misclassified as S (82.18%)	Misclassified as S (97.45%)
2	B	100%	97.01%	Misclassified as W (70.66%)
3	C	100%	99.94%	92.47%
4	D	100%	99.01%	Not detected
5	E	99.77%	Misclassified as S (81.86%)	Misclassified as S (93.87%)
6	F	99.91%	97.52%	79.49%
7	G	100%	100%	86.70%
8	H	100%	100%	94.05%
9	I	100%	99.80%	73.98%
10	J	98.71%	99.67%	80.22%
11	K	100%	99.96%	Misclassified as R (88.59%)
12	L	100%	98.29%	82.97%
13	M	100%	84.41%	Misclassified as S (94.26%)
14	N	100%	95.82%	Misclassified as S (76.98%)
15	O	100%	99.98%	97.74%
16	P	99.99%	99.84%	82.58%
17	Q	100%	99.99%	99.98%
18	R	99.94%	98.98%	89.38%
19	S	99.61%	99.62%	95.84%
20	T	100%	98.12%	88.01%
21	U	99.99%	Misclassified as R (89.99%)	Misclassified as R (74.68%)
22	V	100%	99.03%	77.53%
23	W	100%	100%	99.57%
24	X	100%	99.98%	99.73%
25	Y	100%	97.60%	77.69%
26	Z	99.34%	98.54%	84.22%

Based on Table 3, the system achieved the highest accuracy at a distance of 15 cm, with most letters recognized perfectly. A slight decrease in accuracy was observed for some letters, but it remained above 98%. At 30 cm, accuracy declined, especially for letters A, E, and U, which were frequently misrecognized, although letters like G and W remained

accurate. At 50 cm, accuracy dropped significantly, with some letters not detected and classification errors increasing due to the hand images becoming too small to be recognized properly.

3.4. Testing with Noise

Table 4. Testing with Noisy Audio

No	Condition	Speaker 1	Speaker 2	Notes
1	Quiet environment	35–40 dB	None	Transcription accurate
2	Loud music	65–69 dB	65–69 dB	Sentence did not appear
3	Soft music	65–69 dB	40–45 dB	Sentence appeared completely
4	Loud music with soft speech	40–50 dB	65–59 dB	Sentence did not appear

Speech-to-Text testing was conducted using two speakers: speaker 1 to utter words or sentences, and speaker 2 to play background noise (music). The results showed that the system performed very well in a quiet environment, with accurate transcription. When loud music from speaker 2 overpowered speaker 1's voice, the system failed to recognize the speech. Under relaxed music conditions, the system could still recognize the voice if speaker 1 was dominant. This indicates that the system heavily depends on the quality and clarity of the input audio, and is therefore best used in low-noise environments.

4. CONCLUSION

This study successfully developed a web-based application that facilitates two-way communication between individuals with hearing impairments and the general public. The application translates SIBI (Indonesian Sign System) gestures into text using a CNN model and MediaPipe, and converts speech into text using the Web Speech API, which is then displayed as SIBI letter images.

The gesture recognition feature demonstrated high accuracy, ranging from 98.71% to 100%, supported by training the CNN model with 52,000 hand landmark data points. The speech-to-text feature also performed well, accurately converting speech into text and displaying the corresponding SIBI letter images for both individual letters and full sentences.

Distance testing showed the highest accuracy at 15 cm, decreasing at 30 cm, and dropping significantly at 50 cm due to the reduced clarity of the hand in the camera view. In noisy environments, the system worked optimally in quiet conditions. Loud background noise, such as loud music, caused transcription failures, but in softer ambient noise or when the user's voice was more dominant, the system remained fairly accurate. This indicates that speech-to-text performance is highly dependent on audio quality.

Overall, the integration of gesture recognition and speech-to-text conversion in a single web platform offers an effective and accessible communication solution without requiring additional installation.

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

Future development can focus on expanding gesture detection to words or sentences, improving accuracy under varied lighting and noise conditions, and optimizing the system for mobile devices with features like noise reduction.

REFERENCES

[1] I. Y. Simamora, M. Zahra, W. A. Sinaga, H. E. Pandiangan, and S. F. Hasibuan, "Peran Komunikasi dalam Pembangunan Pendidikan," *J. Pendidik. Tambusai*, vol. Volume 8, p. 8, 2024.

[2] Lela Ayu Septyani, Hanik Noor Solikhah, and Arcivid Chorynia Ruby, "Analisis Penggunaan Bahasa SIBI Untuk Meningkatkan Komunikasi Siswa Tunarungu Dalam Kehidupan Sehari-hari," *J. LENTERA J. Stud. Pendidik.*, vol. 6, no. 2, pp. 135–140, 2024, doi: 10.51518/lentera.v6i2.210.

[3] R. H. Alfikri, M. S. Utomo, H. Februariyanti, and E. Nurwahyudi, "Pembangunan Aplikasi Penerjemah Bahasa Isyarat Dengan Metode Cnn Berbasis Android," *J. Teknoinfo*, vol. 16, no. 2, p. 183, 2022, doi: 10.33365/jti.v16i2.1752.

[4] D. Novaliendry, K. Budayawan, R. Auvi, B. R. Fajri, and Y. Huda, "Design of Sign Language Learning Media Based on Virtual Reality," *Int. J. online Biomed. Eng.*, vol. 19, no. 16, pp. 111–126, 2023, doi: 10.3991/ijoe.v19i16.44671.

[5] G. Amprimo, G. Masi, G. Olmo, and C. Ferraris, "Deep Learning for hand tracking in Parkinson's Disease video-based assessment: Current and future perspectives," *Artif. Intell. Med.*, vol. 154, no. October 2023, 2024, doi: 10.1016/j.artmed.2024.102914.

[6] Purwono, A. Ma'arif, W. Rahmaniar, H. I. K. Fathurrahman, A. Z. K. Frisky, and Q. M. U. Haq, "Understanding of Convolutional Neural Network (CNN): A Review," *Int. J. Robot. Control Syst.*, vol. 2, no. 4, pp. 739–748, 2022, doi: 10.31763/ijrcs.v2i4.888.

[7] M. M. Taye, "Understanding of Machine Learning with Deep Learning :," *Comput. MDPI*, vol. 12, no. 91, pp. 1–26, 2023.

[8] N. Lubis, M. Z. Siambaton, and R. Aulia, "Implementasi Algoritma Deep Learning pada Aplikasi Speech to Text Online dengan Metode Recurrent Neural Network (RNN)," *sudo J. Tek. Inform.*, vol. 3, no. 3, pp. 113–126, 2024, doi: 10.56211/sudo.v3i3.583.

[9] I. Bakti and M. Firdaus, "Arsitektur Convolutional Neural Network InceptionResNet-V2 Untuk Pengelompokan Pneumonia Chest X-Ray," *J. Komput. dan Teknol.*, vol. 01, no. 02, pp. 35–42, 2023.

[10] I. Šušter and T. Ranisavljević, "Optimization of MySQL database," *J. Process Manag. New Technol.*, vol. 11, no. 1–2, pp. 141–151, 2023, doi: 10.5937/jouproman2301141q.

[11] J. W. Kim, J. Y. Choi, E. J. Ha, and J. H. Choi, "Human Pose Estimation Using MediaPipe Pose and Optimization Method Based on a Humanoid Model," *Appl. Sci.*, vol. 13, no. 4, 2023, doi: 10.3390/app13042700.

[12] S. N. Koyineni, G. K. Sai, K. Anvesh, and T. Anjali, "Silent Expressions Unveiled: Deep Learning for British and American Sign Language Detection," *Procedia Comput. Sci.*, vol. 233, pp. 269–278, 2024, doi: 10.1016/j.procs.2024.03.216.

[13] S. Supiyandi, M. Zen, C. Rizal, and M. Eka, "Perancangan Sistem Informasi Desa Tomuan Holbung Menggunakan Metode Waterfall," *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 2, p. 274, 2022, doi: 10.30865/jurikom.v9i2.3986.

[14] J. Gupta, S. Pathak, and G. Kumar, "Deep Learning (CNN) and Transfer Learning: A Review," *J. Phys. Conf. Ser.*, vol. 2273, no. 1, 2022, doi: 10.1088/1742-6596/2273/1/012029.

[15] U. Syach and S. W. M. Edi, "Perancangan Aplikasi Web Manajemen Data Produk Bisnis Perhiasan Berbasis Flask Dan Mongodb," *IT-Explore J. Penerapan Teknol. Inf. dan Komun.*, vol. 3, no. 2, pp. 162–176, 2024, doi: 10.24246/itexplore.v3i2.2024.pp162-176.

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id