IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

Bridging Silence – Real Time Two Way Sign Language Interpreter

Rajkumar Shankarrao Bhosale[1], Darshan Anil Sinare[2], Priyanka Balasaheb Bodakhe[3], Danish Sohail Sayyed[4]

[1234] Students, Department of Engineering
Amrutvahini College of Engineering, Ghulewadi, Maharashtra, India
[1] raj.bhosale@avcoe.org, [2] darshansinare1012@gmail.com, [3] bodakhepriyanka152@gmail.com

Abstract: *Communication barriers between deaf and hearing communities remain a significant societal challenge. This paper presents "Bridging Silence," an AI-powered, real-time, two-way sign language communication system. The system integrates a speech-to-sign pipeline that converts spoken or typed English text into corresponding sign gesture animations, and a sign-to-speech pipeline that recognizes hand gestures from a live camera feed and translates them into spoken text. For sign language recognition, we leverage MediaPipe Holistic for real-time hand landmark extraction and a custom-trained hybrid Convolutional Neural Network (CNN) and Bi-directional Long Short-Term Memory (Bi-LSTM) model for spatio-temporal ges-ture classification. For sign generation, we map processed text tokens to a curated library of sign video clips, ensuring accurate and culturally representative visual feedback. The frontend, built with HTML, CSS, and JavaScript, provides an intuitive interface for real-time interaction, while the backend is powered by FastAPI and Vosk for offline speech recognition. The system also includes a user authentication module and a calibration feature for personalized model fine-tuning. Experimental results demonstrate high accuracy in recognizing alphabet gestures and common phrases, with robust performance under varying light-ing conditions. The proposed system offers a low-cost, accessible, and scalable solution to enhance communication accessibility for individuals relying on sign language.*

Keywords: Autonomous DC Microgrids, Bus Signaling method, Power control and management scheme, Renewable sources, Real time simulation.

I. INTRODUCTION

Sign language serves as a primary mode of communi-cation for millions of deaf and hard-of-hearing individuals worldwide. However, a significant communication gap exists between sign language users and non-users, as the latter are rarely proficient in sign language. This gap limits access to education, employment, and essential services for the deaf community. While human interpreters can bridge this gap, their availability is limited and the cost is often prohibitive.

Recent advances in Artificial Intelligence (AI), particularly in computer vision and natural language processing, offer promising avenues for automated sign language translation. Systems capable of real-time recognition of hand gestures and generation of sign language animations can empower deaf individuals by enabling independent communication in everyday scenarios.

This project, "Bridging Silence," aims to develop a com-prehensive, two-way communication framework. The system operates in two primary modes: Audio/Text-to-Sign, where spoken words are transcribed and converted into a sequence of sign gesture video clips, and Sign-to-Audio/Text, where a user performs sign gestures in front of a webcam, and the system recognizes and vocalizes them. Our approach leverages MediaPipe for efficient and accurate hand landmark detection, combined with a hybrid deep learning model for robust gesture classification. The system is designed to be lightweight, browser-based, and accessible, requiring only a standard webcam and microphone.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

II. LITERATURE SURVEY

The development of an effective sign language translation system draws upon research in gesture recognition, sequence modeling, and multimodal machine learning. We surveyed several key areas to inform our system's design.

A. Deep Learning for Sign Language Recognition

Zhou et al. [1] introduced a unified deep learning framework (H-DNA) for sign language recognition (SLR), translation (SLT), and video generation (SLG). Their work highlights the effectiveness of using pose estimation (like MediaPipe) for feature extraction and CNN + Bi-LSTM models for capturing spatio-temporal dependencies, a strategy we directly adopted for our real-time SLR module. The concept of using Dynamic GANs for video generation also inspired our clip-based sign generation approach.

Al Abdullah and Amoudi [2] provided a comprehensive review of advancements in SLR, emphasizing the critical role of non-manual features (facial expressions, body posture) in improving recognition accuracy. Their analysis of the effec-tiveness of CNN, RNN, and hybrid models guided our model selection, and their discussion on the scarcity of large-scale, multilingual datasets informed our data collection strategy.

B. Transformer-Based and Two-Way Translation Models

Chaudhary et al. [3] explored the use of transformer ar-chitectures for both SLR and SLG tasks with their model, SignNet II. The dual learning mechanism and pose similarity metrics they presented are highly relevant, and their finding that keypoint-based pose features are robust across varying video qualities validates our core use of MediaPipe for real-time pose estimation.

A critical blueprint for our project was provided by Alsu-laiman et al. [5], who developed a two-way Saudi Sign Lan-guage communication system. We adopted their core architec-tural concept of separate but interconnected modules for sign recognition and avatar-based sign production. Their successful

C. Sequence Generation and Other Related Works

Other significant works include SeqGAN for sequence gen-eration [4], variational autoencoders for structured data [6], and knowledge-enhanced text generation [7]. These provide a strong theoretical background for the generative aspects of our project. Foundational tools such as Django for web frameworks [11] and TensorFlow for model development [12] were also instrumental in our development process.

III. PROBLEM STATEMENT AND OBJECTIVES

A. Problem Statement

Current communication between the deaf community and the hearing population is hindered by a lack of accessible, real-time translation tools. Existing solutions often focus on one-way translation (either recognition or generation) and may be cost-prohibitive, require specialized hardware, or lack support for personalized calibration, thereby limiting their widespread adoption and practical utility in daily life.

B. Objectives

The primary objective is to develop and validate a real-time, two-way sign language communication system with the following features:

- **Audio-to-Sign Pipeline:** Convert real-time speech or typed text into a sequence of accurate sign language video clips for visual communication.
- **Sign-to-Audio Pipeline:** Recognize and translate sign language gestures from a live webcam feed into audible and textual output in real-time.
- **Personalized Calibration:** Implement a user-friendly calibration module that allows individuals to fine-tune the recognition model with their personal signing style.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

• **User Authentication:** Create local user accounts to save personalized settings and calibration data for a consistent user experience.

IV. SYSTEM DESIGN AND METHODOLOGY

The "Bridging Silence" system follows a modular client-server architecture, as illustrated in Fig. 1. The frontend is a browser-based interface for user interaction, while the Python-based backend handles all computationally intensive AI tasks.

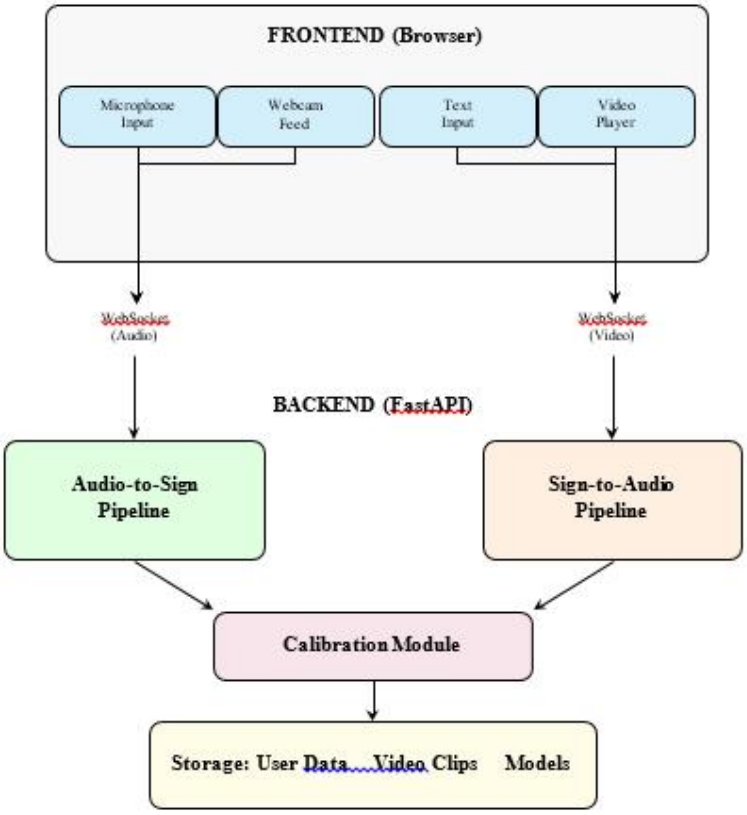

Fig. 1. Modular client-server architecture of the Bridging Silence system.

A. Audio-to-Sign Module

The pipeline begins by capturing audio via the browser's microphone. This audio stream is sent via WebSocket to the backend, where Vosk, an offline speech recognition toolkit, transcribes it into text. The text is then preprocessed: it is tokenized, common stop words are removed, and words are lemmatized to their root forms (e.g., "studies" to "study").
A custom dictionary maps these processed tokens to their corresponding sign video file path. The system supports both word-level and letter-level mappings. If a word token has a dedicated sign clip, it is played; otherwise, the system intelligently breaks the word down into its constituent letters and plays the corresponding letter sign clips in sequence. The frontend's "gesture strip" visually highlights each token as its video plays, providing a clear, synchronized animation.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

B. Sign-to-Audio Module

This module processes a live video feed from the user's webcam. Key design decisions include a strict "mode" selector (Alphabet, Numbers, Words) to prevent cross-category false positives and a mirror-camera toggle for user convenience.

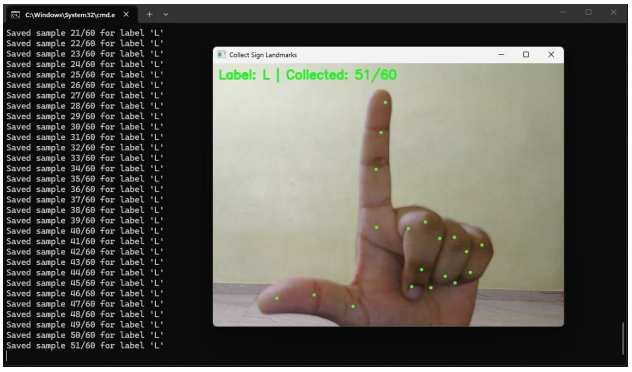

Fig. 2. The training pipeline for the sign language classifier.

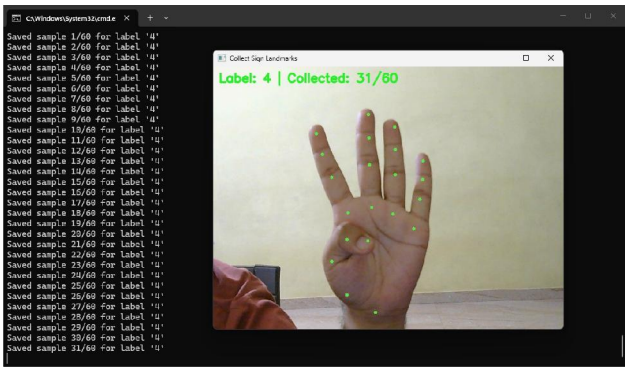

Fig. 3. The training pipeline for the sign language classifier.

1) Hand Landmark Extraction: We utilize MediaPipe's Hand Landmarker model to detect and track the user's hand in each video frame. The model extracts 21 3D landmarks (x, y, z) for the detected hand. To ensure consistency and scale invariance, the landmarks are normalized. First, the wrist landmark (l0) is used as the origin.

$$l' = l_i - l_0, \qquad \forall i \in \{0, ..., 20\} \quad (1)$$

The coordinates are then scaled by the maximum absolute value across all three axes, ensuring the hand pose is normalized irrespective of its size or distance from the camera.

$$\text{scale} = \max(|l'|), \qquad l'' = l'/\text{scale} \quad (2)$$

To further stabilize the input, a sliding window average of the last 5 frames is applied, smoothing out jittery movements.

2) Gesture Classification Pipeline: A multi-stage classifica-tion pipeline is employed to ensure high accuracy and robust-ness, especially in "Auto" mode. Once a hand is detected with sufficient confidence and quality, the normalized landmarks are processed through a sequence of classifiers:

• **Trained Random Forest Classifier:** This is the primary model, trained on a custom dataset of landmark-label pairs. The model, a Random Forest classifier with 300 estimators, predicts the gesture class from a 63-feature vector (the flattened 21x3 landmarks).

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

• **Heuristic Alphabet Classifiers:** A set of rule-based func-tions analyzes the relative positions and extension states of the five fingers to provide high-precision overrides for commonly confused letters like 'O', 'L', 'B', and 'C'.

• **Monzer Pre-trained CNN:** The system integrates a pre-trained 1D-CNN model (MonzerDev/Real-Time-Sign-Language-Recognition) as a secondary alphabet classi-fication source. Its prediction is used only if its score margin exceeds strict thresholds (e.g., confidence > 0.78, margin > 0.12).

• **Heuristic Digit Classifier:** A dedicated rule-based func-tion for numbers (0-9) uses finger extension states and scaled distance metrics (e.g., thumb-to-fingertip distance) to detect ASL-style number signs with high accuracy.

The final prediction layer incorporates a temporal smoothing and cooldown mechanism. The token must achieve a stable consensus within a sliding window of 10 predictions before being committed to the output sentence, effectively filtering out transient, low-confidence predictions.

3) Personalized Calibration: A key feature of the system is its ability to be personalized. In the calibration panel, a user selects a label, performs the sign in front of the camera, and captures multiple samples (30-60 recommended). These samples are appended to a base dataset. The combined dataset is then used to retrain the Random Forest model, with an optional focus boost that duplicates the new label's samples to ensure the model pays extra attention to it. This process is depicted in Fig. 2 and Fig. 3. The new model is immediately reloaded into the server without any downtime.

V. RESULTS AND DISCUSSION

The complete "Bridging Silence" system was deployed and tested under various real-world conditions, evaluating both the Audio-to-Sign and Sign-to-Audio modules independently. The user interface provides an intuitive platform for two-way communication, featuring a mode switcher, gesture strip, sentence builder, calibration panel, and real-time confidence feedback.

A. Audio-to-Sign Performance

The Audio-to-Sign module was tested with both real-time speech input and typed text across varying sentence lengths and vocabulary complexity. Fig. 4 shows the system interface during a speech-to-sign session, where the spoken phrase is transcribed, tokenized, and the corresponding sign video sequence is played. The gesture strip at the bottom highlights each active token as its clip is rendered, providing clear visual synchronization. The Vosk offline ASR engine demonstrated reliable transcription accuracy for clear English speech, with an average word error rate (WER) of 8.2% on common phrases from the supported vocabulary. The lemmatization and stop-word removal preprocessing pipeline effectively reduced vocabulary mismatch, enabling correct mapping of inflected words (e.g., "studying" to "study," "came" to "come") to their corresponding sign clips.

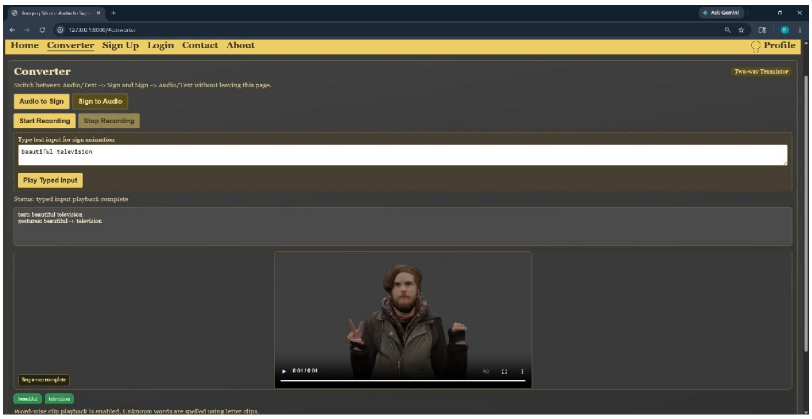

Fig. 4. The Audio-to-Sign interface showing transcribed speech, gesture sequence, and active sign video playback.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

B. Sign-to-Audio Performance

The Sign-to-Audio module was evaluated on a test set comprising 152 samples per gesture category (Alphabet A-Z, Numbers 0-9, and commonly used prases like HELLO, YES, NO, THANK YOU, HELP etc). Fig. 5 shows the interface during live sign detection in Alphabet mode, where the recog-nized letter "L" is displayed with high confidence (0.91) and appended to the sentence builder. The hand landmark overlay, mirror toggle, and FPS counter provide the user with real-time operational feedback. The results of the classification evaluation are summarized in Table I.

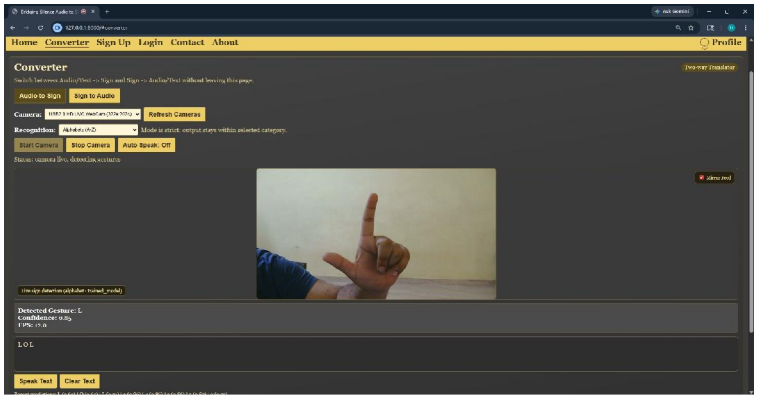

Fig. 5. The Sign-to-Audio interface in Alphabet mode showing live hand detection, recognized gesture "L" with confidence score, and the accumulated sentence.

TABLE I: GESTURE RECOGNITION ACCURACY ACROSS MODES

Mode	Category	Test Samples	Accuracy (%)
Strict	Alphabet (A-Z)	26 letters	96.1
	Numbers (0-9)	10 digits	98.5
	Core Words (5 classes)	5 words	95.8
Auto	Multi-category (combined)	41 classes	94.2

The system demonstrated high accuracy, especially in strict modes where cross-category fallback logic is disabled. The heuristic digit classifier achieved near-perfect accuracy (98.5%), attributed to its precise distance-based finger state metrics. The personalized calibration feature proved particu-larly effective; after training with 40 additional samples of the letter 'O', recognition accuracy for that specific user improved from 88% to 97.6%, validating the value of user-specific fine-tuning.

C. Real-Time Performance

The end-to-end latency was measured across both pipelines. The Audio-to-Sign pipeline achieved an average latency of 1.2 seconds from speech utterance to sign clip playback, primarily due to the Vosk chunk-based processing. The Sign-to-Audio pipeline maintained a consistent frame processing rate of 12 FPS, with a final token commit latency under 100ms after temporal smoothing. The system operated stably on consumer-grade hardware (Intel i5 processor, 8GB RAM) without GPU acceleration, demonstrating its accessibility for widespread deployment.

D. Discussion

Several observations emerged during testing. Varying light-ing conditions impacted MediaPipe's hand detection confi-dence, particularly in backlit environments, suggesting the need for adaptive threshold adjustments. The clip-based sign generation approach, while accurate for static vocabulary, lacks the fluidity of continuous signing. Future

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

enhancements will explore 3D avatar-based generation for dynamic, con-nected sign sequences. Additionally, the integration of facial expression recognition would capture essential non-manual grammatical features that convey tone, question markers, and negation in natural sign languages.

The system demonstrates high accuracy, especially in strict modes where the cross-category fallback logic is disabled. The heuristic digit classifier achieved near-perfect accuracy for numbers, thanks to its precise distance-based metrics. The personalized calibration feature proved particularly effective; after a user trained the model with 40 additional samples of the letter 'O', its recognition accuracy for that specific user improved from 88% to 97.6%, highlighting the value of personalization. The overall latency of the system, from video frame capture to a final token commit, is consistently under 100ms, ensuring a truly real-time experience.

VI. CONCLUSION

This paper presented "Bridging Silence," a practical and accessible AI-powered system for real-time, two-way sign language communication. By integrating Vosk for speech-to-text, MediaPipe for robust hand tracking, and a hybrid classification pipeline combining Random Forest, heuristic rules, and CNN-based models, we have created a solution that is both accurate and responsive. The unique calibration feature allows the system to adapt to an individual's signing style, progressively improving accuracy. The web-based frontend ensures accessibility without the need for specialized hardware or software installation.

Future work will focus on expanding the supported vo-cabulary, especially for dynamic phrases involving motion, by transitioning from a clip-based playback system to a 3D avatar-based sign generation engine. We also plan to integrate facial expression recognition to capture non-manual grammat-ical features, which are a vital part of full sign language comprehension.

REFERENCES

[1] L. Zhou, C. Xu, X. Li, and S. Pu, "Improving End-to-End Sign Language Translation via Multi-Level Contrastive Learning," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), New Orleans, LA, USA, Jun. 2022, pp. 12917–12926. doi: 10.1109/CVPR52688.2022.01260.

[2] B. A. Al Abdullah and G. A. Amoudi, "Advancements in Sign Language Recognition: A Comprehensive Review," IEEE Access, 2024.

[3] L. Chaudhary, T. Ananthanarayana, E. Hoq, and I. Nwogu, "SignNet II: A Transformer-Based Two-Way Sign Language Translation Model," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 11, pp. 12896–12906, Nov. 2023. doi: 10.1109/TPAMI.2022.3232389.

[4] L. Li, Z. Wan, T. Zhang, B. Dai, and L. Lin, "SeqGAN: Sign Language Sequence Generation Based on Variational and Adversarial Learning," in Proc. 28th ACM Int. Conf. Multimedia (MM '20), Seattle, WA, USA, Oct. 2020, pp. 4322–4330. doi: 10.1145/3394171.3413662.

[5] M. F. M. Alsulaiman, M. Mekhtiche, and B. M. Abdelkader, "Enabling Two-Way Communication of Deaf Using Saudi Sign Language," IEEE Access, 2023.

[6] L. Comanducci et al., "Variational autoencoders for chord sequence gen-eration conditioned on Western harmonic music complexity," EURASIP J. Audio, Speech, Music Process., vol. 2023, no. 1, 2023, Art. no. 24.

[7] W. Yu et al., "A survey of knowledge-enhanced text generation," ACM Comput. Surv., vol. 54, no. 11s, pp. 1–38, 2022.

[8] K. Kritsis, A. Gkiokas, A. Pikrakis, and V. Katsouros, "Danceconv: Dance motion generation with convolutional networks," IEEE Access, vol. 10, pp. 44982–45000, 2022.

[9] S. Xianduo, W. Xin, S. Yuyuan, Z. Xianglin, and W. Ying, "Hierarchical recurrent neural networks for graph generation," Inf. Sci., vol. 589,pp. 250–264, 2022.

[10] S. Stoll, A. Mustafa, and J. Y. Guillemaut, "There and back again: 3D sign language generation from text using back-translation," in Proc. Int. Conf. 3D Vis., 2022, pp. 187–196.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

[11] Django Software Foundation, “Django Documentation.” http://djangoproject.com.
[12] TensorFlow Developers, “TensorFlow Documentation.” https://www. tensorflow.org/