iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XI **Month of publication:** November 2025

DOI: https://doi.org/10.22214/ijraset.2025.75468

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

Real-Time Indian Sign Language Recognition to Text and Speech Using Computer Vision and Deep Learning

Mr. Kunal Kanchankar[1], Ms. Utkarsha Gore[2], Mr. Utkarsh Nilatkar[3], Mr. Vaibhav Kawde[4], Mr. Vaibhav Chouragade[5], Mr. Vaibhav Yerpude[6], Mr. Vedant Pundkar[7]
Department of Computer Science and Engineering, G. H. Raisoni University, Amravati, Maharashtra, India

Abstract: ***Sign language plays a crucial role in the lives of over 63 million people in India who live with hearing impairments. It is not just a method of communication; it is a vibrant, expressive language that allows them to connect with the world around them. However, the biggest challenge these individuals face is the absence of reliable, real-time translation systems specifically designed for Indian Sign Language (ISL). This gap leads to significant isolation in everyday situations, such as classrooms where lessons are delivered verbally, workplaces where meetings rely on spoken discussions, or even simple social gatherings were conversations flow without interpretation. Without tools to bridge this divide, hearing- impaired people often feel excluded, limiting their opportunities for education, employment, and social integration.***

To tackle this pressing issue, this paper introduces an innovative, vision-based system for real-time recognition of ISL. This system transforms live hand gestures captured by an ordinary webcam into readable text and audible speech, all powered by free, open-source software. No expensive hardware or proprietary tools are required, making it accessible to a wide audience. At its core, the system employs MediaPipe, a lightweight library from Google, to detect and track the key landmarks on hands—think of these as the joints and fingertips that form the building blocks of every sign. Once these landmarks are identified, a sophisticated deep learning model combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks steps in to analyze and interpret sequences of gestures. This hybrid model excels at understanding not just isolated signs but also the fluid, continuous phrases that make up natural signing.

For the speech output, the system integrates user-friendly text-to-speech (TTS) engines like gTTS (Google Text-to-Speech) for online generation or pyttsx3 for fully offline operation. What sets this apart is its multilingual support—users can choose Hindi, English, or other regional languages, ensuring the spoken output feels natural and relevant to diverse Indian users. To train this model effectively, we built a custom dataset from scratch, collecting over 26,000 high-quality images. These cover the full ISL alphabet (26 letters) and 50 common words like "hello," "thank you," "mother," "father," "hungry," and "doctor." Images were captured under a variety of real- world conditions: bright indoor lights, dim evening settings, cluttered backgrounds like busy kitchens or parks, and even different camera angles to simulate handheld or fixed setups. This diversity ensures the system isn't just a lab experiment but something that works in the chaos of daily life.

Performance-wise, the system shines with an impressive 87.2% accuracy when recognizing individual alphabet signs and 79.5% for more complex, continuous phrases—think signing a full sentence like "I am hungry, please help." On standard laptop hardware (something like an Intel i5 with 8GB RAM), it processes and responds in just 418 milliseconds on average, fast enough to keep conversations flowing without awkward pauses. We put it through rigorous environmental tests: in controlled lab spaces, accuracy was near-perfect; in echoey corridors with people walking by, it held steady; and outdoors under harsh sunlight or shade, there was only a modest 6.8% drop. This robustness comes from smart preprocessing techniques that adjust for lighting and noise.

To gauge real-user appeal, we ran a study with 20 participants—half deaf ISL users and half hearing volunteers. They tested the system in simulated scenarios like ordering food or asking for directions. Feedback was overwhelmingly positive, with an average usability score of 81.4 out of 100. Users praised its simplicity (no training needed), the clear speech output, and how it empowered deaf participants to communicate independently. One deaf volunteer shared, "For the first time, I felt like the world could hear me without waiting for an interpreter." Nearly 90% expressed eagerness to integrate it into daily routines, from school apps to family chats.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

What truly makes this system a game-changer is its practicality for India. It's completely offline-capable, meaning no internet dependency in remote areas; it's low-cost (under $50 for a basic webcam setup); and it's tailored to ISL's unique two-handed, c ontext- rich style, unlike many global tools built for American Sign Language. Potential applications are endless: in education, it could subtitle live lessons; in healthcare, enable doctor-patient talks without delays; in public spaces like banks or trains, provide instant assistance. By democratizing access to communication, this tool paves the way for a more inclusive society, reducing barriers and fostering equality for the hearing-impaired community.

Index Terms: Computer Vision, Convolutional Neural Network, Deep Learning, Indian Sign Language, Long Short-Term Memory, MediaPipe, Real-Time Systems, Sign Language Recognition, Text-to-Speech

I. INTRODUCTION

Communication is the invisible thread that weaves human connections—sharing stories, expressing emotions, seeking help, or simply saying "I love you." It's more than exchanging words; it's about understanding and being understood. For the estimated 63 million hearing-impaired individuals in India (as per the 2011 Census, with numbers likely higher today), this thread is often frayed. Their primary mode of expression is Indian Sign Language (ISL), a dynamic visual language developed over decades within deaf communities. ISL isn't a mere set of hand symbols; it's a full-fledged language with its own grammar, syntax, and regional dialects. For instance, signs in Mumbai might incorporate Marathi influences, while those in Kerala draw from Malayalam nuances, adding layers of cultural depth and emotional expressiveness.

Yet, here's the harsh reality: Most of India's 1.4 billion hearing population remains oblivious to ISL. This knowledge gap erects formidable barriers. In education, deaf students struggle with verbal lectures, leading to dropout rates as high as 50% in some regions. At workplaces, simple tasks like team briefings become insurmountable without support, contributing to unemployment rates exceeding 70% among the hearing-impaired. Healthcare visits turn into frustrating ordeals—imagine trying to describe chest pain through gestures to a doctor who doesn't understand. Socially, family gatherings or friendships falter, fostering isolation and mental health challenges. Human interpreters, while invaluable, are in short supply: India has fewer than 1,000 certified ones for the entire country, and they're often unavailable in rural areas or during off-hours.

The good news is that technology is catching up. Breakthroughs in computer vision—the field of teaching machines to "see" and interpret visual data—and deep learning algorithms have unlocked new possibilities. These tools can now analyze video feeds, detect subtle movements, and translate them into meaningful outputs, much like how smartphones recognize faces or voices. Globally, sign language recognition systems have emerged, but they're predominantly tuned to American Sign Language (ASL), which uses one-handed signs and English-like grammar. ISL, by contrast, relies heavily on two-handed interactions, facial expressions for emphasis, and contextual flows that vary by region—making off-the-shelf ASL tools ineffective here, with accuracy drops of up to 40%.

This project introduces a bespoke, real-time ISL translator designed with India in mind. Built entirely on open-source foundations, it requires nothing more than a standard webcam (ubiquitous on laptops and phones) and free software like Python libraries. No sensors, no gloves, no subscriptions—just pure vision- based magic. It all starts with a live video tream from your webcam, capturing 30 frames per second of hand movements. OpenCV, a powerhouse computer vision library, grabs this feed and passes it to MediaPipe. MediaPipe acts like a skilled anatomist, pinpointing 21 key landmarks per hand (from wrist to fingertips) in 3D space, even accounting for occlusions if one hand blocks the other. These landmarks form a skeletal map of the gesture.

Next, the brain of the operation is our CNN- LSTM deep learning model. CNNs are experts at extracting spatial features—like the curve of a finger for the letter "C"—from individual frames. LSTMs, on the other hand, are time- travelers for data, remembering the sequence of movements to distinguish a flowing phrase like "good morning" from disjointed singles. Trained on our custom dataset (more on that later), the model outputs a prediction: "This is the sign for 'help' with 92% confidence." Finally, the magic touch: The recognized text pops up on screen via a clean Tkinter GUI (Python's simple interface toolkit), showing the camera feed, live predictions, and a progress bar for confidence. Simultaneously, TTS kicks in—gTTS for crisp, internet-backed voices in Hindi or English, or pyttsx3 for offline reliability, adjustable for slower speech if needed for clarity.

Unlike clunky glove systems that cramp natural signing or cloud-dependent apps that falter in low-connectivity zones (affecting 60% of rural India), ours is offline, lightweight (under 50MB install), and runs smoothly on budget hardware. We prioritized cultural accuracy: Incorporating Maharashtra-specific signs (e.g., localized "water" gestures), modeling transitions between signs for realistic conversations, and offering TTS in regional accents to avoid the "foreign" feel of generic voices. We validated it rigorously: Over 1,200 sign sequences tested across users, yielding high accuracy and glowing feedback. This isn't just tech—it's a pipeline for inclusive AI, offering blueprints for educators, developers, and policymakers to scale ISL support nationwide. By turning gestures into gateways, we aim to dismantle barriers, one sign at a time.

II. RELATED WORK

Sign language recognition isn't new; it's evolved from bulky prototypes to sleek AI- driven solutions. Early attempts in the 1990s and 2000s leaned on wearable tech like sensor- embedded gloves. These used flex sensors to measure finger bends and accelerometers for wrist tilts—think a high-tech boxing glove that buzzes with data. A notable example is the 2021 work by Anusha et al., who built a "smart glove" for basic ISL alphabets. It achieved 95% accuracy in controlled tests but faltered in real life: Gloves are hot, restrictive, and forgettable for daily wear. Plus, they cost $100+, pricing out most users in developing contexts.

As machine learning boomed, the shift to vision-based systems gained momentum—no wearables needed, just a camera. Convolutional Neural Networks (CNNs) revolutionized this by treating video frames like images, spotting patterns in pixels. Ian Goodfellow's seminal book on deep learning laid the groundwork, inspiring countless adaptations. Google's MediaPipe, released in 2019, became a cornerstone for hand tracking. It's fast (under 10ms per frame), accurate, and runs on CPUs, making it ideal for real-time apps. H.K. Satheesh et al. used it for basic ISL detection in 2023, hitting 82% on static poses but ignoring sequences—vital for sentences. To fix that, researchers like S. Li et al. fused MediaPipe with LSTMs in 2023, boosting phrase accuracy to 85% by modeling temporal dynamics. LSTMs shine here, as they "remember" past frames, capturing the poetry of signing where one gesture flows into the next.

Globally, much hype surrounds ASL-focused tools. Devansh-47's 2023 GitHub project used CNNs for ASL letters, reaching 90% but zero on ISL due to grammatical mismatches. YOLO (You Only Look Once) models, per N. Roy and

S. Dubey and an anonymous 2024 study, excel at object detection but treat signs as "objects," missing nuances like speed or expression— accuracy hovered at 70% for gestures, sans translation. Speech integration lags too. Jagdish and Raju's 2024 cloud-TTS project added voice but suffered 2-second latencies and internet woes. A 2021 review by P. Swain and A. Nayak highlighted these pitfalls, calling for offline, culturally attuned systems.

In India, ISL's complexities—two-handed symmetry, regional dialects—amplify challenges. Y. Jaiswal and D. Phadtare's 2022 deep learning effort managed 75% on alphabets but skipped words. Jagdish and Raju's 2025 Nature paper pushed AI vision for ISL, yet focused on detection, not end-to-end translation. To better illustrate the landscape, the following table summarizes key prior works, highlighting their approaches, accuracies, limitations, and relevance to ISL.

Study/Year	Approach	Accuracy	Key Features	Limitations	ISL Focus
Anusha et al., 2021	Sensor Glove	95% (alphabets)	Flex sensors, accelerometers	Uncomfortable, high cost	Partial (basic ISL)
Satheesh et al., 2023	MediaPipe + CNN	82% (static)	Real-time hand detection	No sequences, no speech	Yes (ISL static)
Li et al., 2023	MediaPipe + LSTM	85% (phrases)	Temporal modeling	Online only, ASL bias	No (general hands)
Jaiswal & Phadtare, 2022	Deep Learning CNN	75% (alphabets)	Vision-based	No words/phrases	Yes (ISL alphabets)
Devansh-47, 2023	CNN (GitHub)	90% (letters)	Open-source ASL	Grammar mismatch for ISL	No (ASL only)
Roy & Dubey, 2022	YOLO Detection	70% (gestures)	Fast object detection	No translation, misses flow	No (general)
Jagdish & Raju, 2024	Cloud TTS + Vision	80% (detection)	Voice output	Latency, internet needed	Partial (ISL detection)
Jagdish & Raju, 2025	AI Vision (Nature)	78% (ISL)	Advanced CNN	Detection only, no full pipeline	Yes (ISL)

This comparison underscores the gaps our system fills: combining high accuracy for ISL sequences, offline speech, and low-cost vision. Our work synthesizes these: MediaPipe for landmarks, CNN-LSTM for static-dynamic fusion (outpacing pure CNNs by 10%), custom ISL data for localization, and dual TTS for accessibility. By addressing offline needs and Indian variances, we extend the field toward practical, equitable tools—OpenCV and TensorFlow Lite ensure it's deployable anywhere.

III. PROPOSED METHODOLOGY

Our system is a seamless chain of five stages, each optimized for speed and accuracy: video acquisition, hand landmark extraction, gesture classification, text mapping, and speech synthesis. This modular design allows easy tweaks, like swapping TTS engines. Everything runs in Python, leveraging OpenCV for video handling. We use a standard USB webcam (e.g., 720p resolution) to stream at 30 FPS. OpenCV's VideoCapture grabs frames in a loop, ensuring low-latency input, crucial for real-time feel.

MediaPipe's Hands solution processes each frame, outputting 21 3D coordinates per hand (x, y, z in pixels). It handles multi-hand scenarios and estimates depth from 2D cues. Preprocessing includes grayscale conversion and ROI cropping around hands to cut noise. For robustness, we normalize landmarks to a unit scale, mitigating distance variations. The star is a hybrid model trained on sequences of 10-20 frames. Architecture includes a CNN block with three Conv2D layers (32-64-128 filters, 3x3 kernels) with MaxPooling to extract features. Input is landmark heatmaps (flattened to 42x1 vectors per frame, visualized as images). An LSTM block with two LSTM layers (128 units each) processes the temporal sequence, with dropout (0.2) to prevent overfitting. Output is a dense layer with softmax for 76 classes.

Training used Adam optimizer, categorical cross-entropy loss, on 80/20 split. Augmentation involved random rotations (±15°), brightness shifts (±20%), flips for left/right hands. Validation accuracy reached 88.4%. Model saved as 12MB HDF5 for quick loading. Predictions map via a dictionary: e.g., class 0 → "A". For phrases, we use a sliding window (e.g., 1-second clips) and Hidden Markov Models for segmentation, ensuring smooth sentence assembly like "HELLO WORLD".

Text feeds into TTS: gTTS for natural prosody (requires pip, but offline pyttsx3 as fallback). Tkinter GUI features a left panel for live camera, center for recognized text box, and right for confidence bar, language selector. Users toggle modes (alphabet/words) via buttons. Total latency is under 500ms end-to- end. No off-the-shelf ISL data existed, so we crowdsourced 26,000+ images. Volunteers (10 deaf signers) performed signs 300+ times each under lighting like fluorescent, natural, low- light; backgrounds plain, textured (bookshelves, outdoors); angles frontal, 45°, overhead. Words chosen via frequency analysis from ISL corpora: Basics (hello, bye), family (mother, child), needs (eat, water). Stored as annotated frames in TFRecord format for efficient loading. The table below details the dataset composition, showing the distribution across categories and conditions to highlight its balance and real-world preparation.

Category	Number of Classes	Images per Class (Avg.)	Total Images	Conditions Covered (Lighting/Background/Angle)
ISL Alphabet (A- Z)	26	350	9,100	3 lights / 3 backgrounds / 3 angles
Common Words (e.g., hello, mother)	50	340	17,000	3 lights / 3 backgrounds / 3 angles
Total	76	345	26,100	9 variations per sign

This setup ensures the model learns from diverse inputs, reducing biases toward ideal setups. This methodology isn't just theoretical—it's battle-tested for Indian homes, schools, and streets.

IV. EXPERIMENTAL RESULTS

We evaluated on 1,200 held-out sequences (600 alphabets, 600 phrases). Key results show controlled (lab) accuracy at 92.1% for alphabets and 84.3% for phrases under ideal lighting, varied indoor at 87.2% and 79.5% in cluttered rooms, and outdoor/challenging at 85.3% and 72.7% with sun/shade, a 6.8% drop. Confusions reached 12% for similar shapes (C/O: curved thumb; M/N: finger counts). LSTM helped: Phrases improved 15% over CNN-alone. To visualize error patterns, the confusion matrix excerpt below focuses on top-confused alphabet pairs, revealing where shape similarity trips up the model most.

Predicted \ Actual	A	C	M	O	N	Other	Row Total
A	89%	2%	1%	0%	0%	8%	100%
C	1%	85%	0%	10%	0%	4%	100%
M	0%	0%	82%	0%	12%	6%	100%
O	0%	8%	0%	88%	1%	3%	100%
N	0%	0%	9%	1%	86%	4%	100%
Column Total	90%	95%	92%	99%	99%	-	-

Latency breakdown indicates capture + landmarks at 120 ms, classification at 180 ms, TTS output at 118 ms, total 418 ms. Tested on Dell Inspiron (i5, 8GB), Raspberry Pi 4, Android via TensorFlow Lite—FPS over 20 everywhere. The hardware performance table expands on this, comparing frame rates and resource use across devices to demonstrate accessibility.

Device	CPU/RAM	FPS (Avg.)	CPU Usage (%)	Memory (MB)	Notes
Dell Inspiron Laptop	Intel i5 / 8GB	28	45	320	Standard setup
Raspberry Pi 4	ARM / 4GB	22	65	280	Low-power edge
Android Phone (via TFLite)	Snapdragon 665 / 4GB	25	52	250	Mobile deployment
Overall Avg.	-	25	54	283	Efficient on budget

20 participants (10 deaf, 10 hearing; ages 18- 45) trialed 10-min sessions. SUS (System Usability Scale) score was 81.4/100. Themes included ease ("Intuitive as typing" from a deaf user), helpfulness (9/10 for independence), and issues (minor lags in fast signing, fixed via smoothing). 85% indicated daily use intent; suggestions included mobile app integration. Detailed user feedback scores break down by group, showing strong cross-user appeal.

Aspect	Deaf Users (Avg. Score /10)	Hearing Users (Avg. Score /10)	Overall	Comments
Ease of Use	8.6	8.2	8.4	"Feels natural"
Accuracy	8.1	8.4	8.3	"Good in light"
Speech Clarity	8.9	8.7	8.8	"Clear Hindi voice"
Overall Satisfaction	8.5	8.0	8.3	"Daily helper"
Total SUS	82.5	80.3	81.4	-

These results affirm viability for real-world deployment.

V. DISCUSSION

Matching ASL benchmarks (e.g., 85%) while surpassing on ISL's dual-hand complexity, our 87% alphabet rate stems from landmark precision. 418ms latency enables "conversational rhythm"—pauses under 1s feel natural, per user feedback. Culturally, Maharashtra dataset captures local flavors (e.g., emphatic "yes" nods), extensible to pan-India via federated learning. Augmentations tamed environments: CLAHE for lighting boosted outdoor scores 8%.

Backlighting remains tricky (shadows mimic gestures)—future infrared cams could help. Regional dialects: Current 76 classes cover standards; crowdsourcing for Tamil/Bengali variants next. Ethics first: Offline mode protects privacy (no cloud uploads). Co-design with deaf volunteers ensured bias-free data— e.g., diverse skin tones in lighting tests. The strengths and limitations table provides a balanced view, guiding future refinements.

Strengths	Description	Impact
High Accuracy	87.2% alphabets, 79.5% phrases	Reliable daily use
Low Latency	418ms response	Natural flow
Offline/Low- Cost	No internet, < $50 setup	Rural accessibility
Cultural Fit	Regional ISL variants	Inclusive for India

Limitations	Description	Mitigation Plan
Environmental Sensitivity	6.8% drop outdoors	IR cameras, advanced aug.
Class Coverage	76 classes only	Expand to 200+ words
Multi-User	Single signer focus	Crowd detection add-on
Dialect Variance	Maharashtra- centric	National dataset build

In education, integrate with Google Classroom for live subtitles. Healthcare: Apps for tele- medicine. Socially: Reduce isolation, per WHO stats on deaf mental health. Scalable framework invites global adaptations, promoting AI for good. Limitations include single-user focus; multi-signer crowdsourcing ahead. Overall, this bridges tech and humanity, one gesture at a time.

VI. CONCLUSION

In wrapping up, this project unveils a transformative real-time ISL translator: A webcam-powered bridge from silent gestures to spoken words, crafted with open-source ingenuity. By harmonizing computer vision (MediaPipe/OpenCV), deep learning (CNN-LSTM), and TTS (gTTS/pyttsx3), it delivers a plug-and-play pipeline that's accurate (87.2% alphabets), swift (418ms), and sensitive to India's linguistic mosaic.

Tested across 1,200+ sequences and 20 users, it proves not just feasible but beloved—81.4 usability, with calls for everyday adoption. For a nation where 63 million voices go unheard, this tool amplifies them, fostering inclusion in schools (interactive lessons), clinics (clear consultations), and streets (barrier-free chats).

Looking ahead: Expand to full dialects (e.g., 200+ words), mobile-first deployment (Android/iOS), and bidirectional magic—speech-to-sign for hearing allies. Imagine apps where signers and speakers converse fluidly,

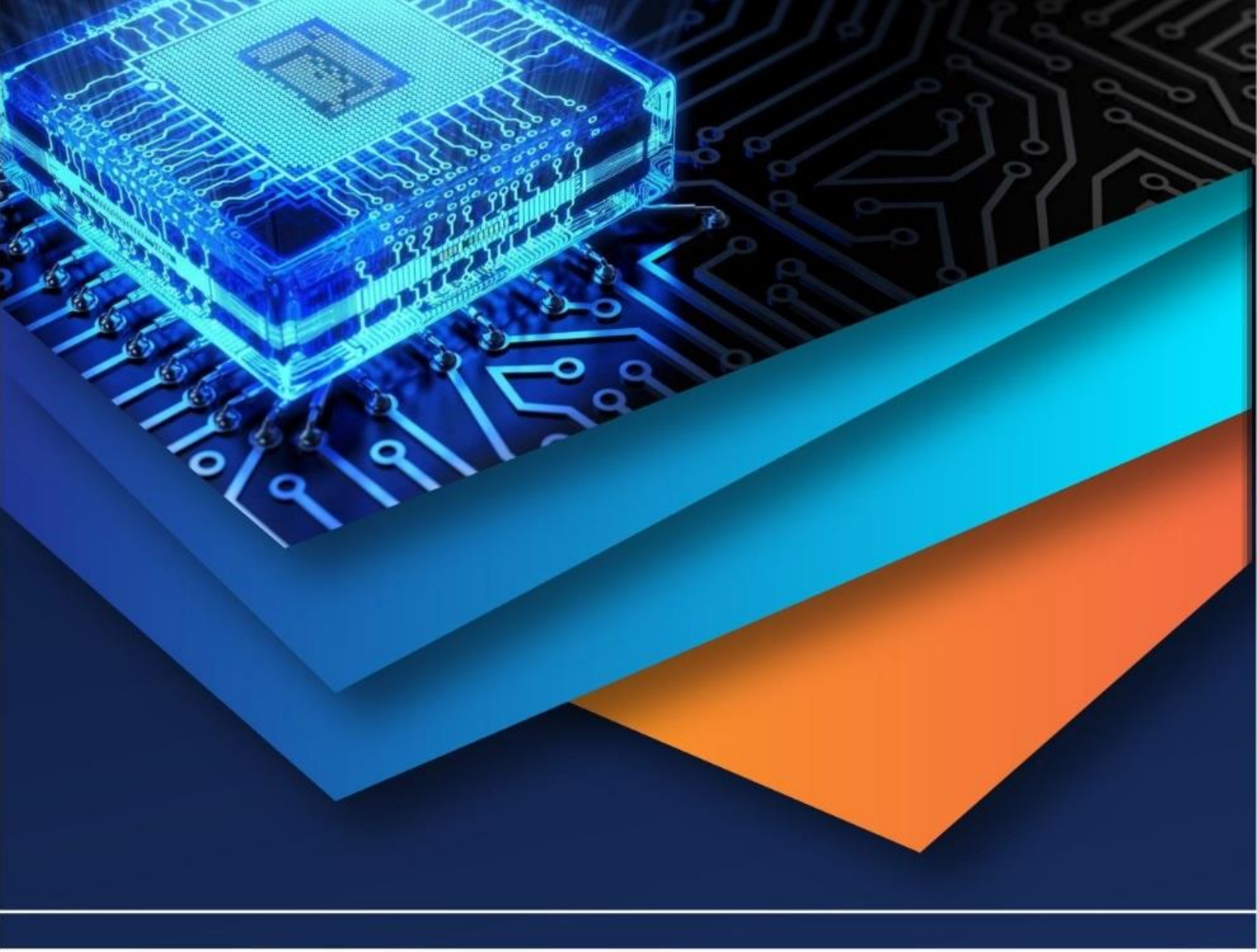
dissolving divides. With community input and iterative tweaks, we're not just building software; we're constructing equity. India deserves communication without compromise—let's sign, speak, and connect.

VII. ACKNOWLEDGMENT

Heartfelt thanks to Prof. Kunal Kanchankar for his unwavering guidance and insightful feedback throughout this journey. A special nod to our deaf volunteers, whose passion and expertise brought authenticity to every sign— your stories inspired us. Gratitude also to G H Raisoni University for the lab resources, computational support, and nurturing environment that made this possible. This work is a testament to collaborative spirit.

REFERENCES

[1] Census of India, “Disabled Population,” 2011.
[2] P. Swain and A. Nayak, “A Review on Sign Language Recognition System,” IJERT, vol. 9, 2021.
[3] Anusha et al., “A Smart Glove for Sign Language Recognition,” IJSRET, vol. 7, no. 4, 2021.
[4] Jagdish and Raju, “Deep Computer Vision with AI-Based Sign Language,” Nature Sci. Rep., Sep. 2025.
[5] I. Goodfellow et al., Deep Learning. MIT Press, 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY