International Journal on Engineering Technology (InJET)

www.kec.edu.np/kec-journal/
Volume 3, issue no. 1, Dec 2025, Page 54 - 61
Received Date: Jul 30, 2025
Accepted Date: Sep 15, 2025
ISSN: 3021-940X (Print)
DOI: https://doi.org/10.3126/injet.v3i1.87013

Development of an Audio to Sign Language Translator using a Random Forest Classifier

Sujit Adhikari[1], Sarishma Neupane[2*], Ranjit Adhiikari [3*], Prayash Niraula[4*], Subash Panday[5*]

[1]Department of Electronics and Computer Engineering, National College of Engineering, Talchhikhel, Lalitpur, Nepal, nce078bct044@nce.edu.np

[1]Department of Electronics and Computer Engineering, National College of Engineering, Talchhikhel, Lalitpur, Nepal, nce078bct038@nce.edu.np

[3]Department of Electronics and Computer Engineering, National College of Engineering, Talchhikhel, Lalitpur, Nepal, nce078bct031@nce.edu.np

[4]Department of Electronics and Computer Engineering, National College of Engineering, Talchhikhel, Lalitpur, Nepal, nce078bct029@nce.edu.np

[5]Department of Electronics and Computer Engineering, National College of Engineering, Talchhikhel, Lalitpur, Nepal, subash@nce.edu.np

Abstract

The ubiquity of English as a global language underscores the necessity for accessible communication solutions that accommodate diverse linguistic and auditory needs. In this context, the paper presents a novel English-to-American Sign Language (ASL) translation system that leverages Natural Language Processing (NLP) and Machine Learning (ML) to convert English text or speech into animated sign gestures. This system applies the Porter Stemming Algorithm to reduce words to their root forms and removes stop words to improve clarity. Word2Vec embedding were employed to transform the pre-processed text into vector representations, which were subsequently classified using a Random Forest model trained on a self-curates ASL dataset consisting of 126 videos, encompassing 90 words,26 alphabets and 10 numbers. The model achieves an accuracy of 94.51%, effectively recognizing base words and their synonyms. For out-of-vocabulary terms, the system defaults to letter-by-letter ASL finger spelling. Developed with Blender for 3D gesture animation and Django for backend processing, the solution offers a scalable and cost-effective model for real-time sign language interpretation.

Keywords: Porter, Natural Language Processing, Blender, Sign language, Machine learning

1. Introduction

Human communication systems have undergone significant transformation over centuries, yet millions of individuals with auditory disabilities remain excluded from mainstream conversational model. (WHO, 2025) Sign languages, particularly American Sign Language (ASL), serve as a vital linguistic tool for the deaf and hard-of-hearing community, providing a structured visual language that mirrors spoken language in function, though not in form. The ability to computationally interpret and generate sign language has become a central concern within the field of assistive technology and human-computer interaction, where the integration of artificial intelligence (AI) and linguistic modelling is increasingly being explored as a medium for inclusive design. (Danielle Bragg, 2019)

Contemporary approaches to sign language translation focus on the identification of gestures through images and the application of convolutional neural networks. (Necati Cihan Camgoz, 2018) Although these systems work effectively in identifying static or video-recorded signs, the more complex task of translating text or speech into meaningful and visually coherent sign gestures is still in its infancy. (Jie Huang, 2018) The reverse translation process requires a deep understanding of linguistic morphology, synonym resolution and the complex mapping of gestures, all of which can be effectively addressed using data-driven models.

In this context, this study proposes a computational framework for English-to-ASL translation that integrates techniques across multiple domains. The initial language normalization leverages Porter's stemming algorithm (Brian Dickinson, 2015) a widely used rule-based approach in information retrieval and text simplification. Word semantics are modelled using Word2Vec embedding, known for their capacity to

* Corresponding author

preserve contextual similarity in vector space . For gesture classification, the system adopts the Random Forest Classifier, which has shown strong generalization in linguistic categorization tasks due to its ensemble architecture. (BREIMAN, 2001)Finally, the output is rendered via Blender, an open-source 3D animation engine previously employed in sign language visualization systems (Blender, 2023) and managed within a Django-based interface for real-time interaction. (Django, 2023)

In essence, this research aims to contribute to the growing body of assistive communication technologies by presenting a scalable and interactive English-to-sign language translation framework. This framework is designed for real-world use in hospitals, schools, and workplaces, with the goal of promoting inclusivity and equal participation in society.

2. Literature Review

Recognition and translation of sign languages into text using convolutional neural networks have shown promising results in recent studies. One such system achieved a recognition accuracy of 93.27%, with a significant reduction in computational time, by utilizing pre-processing and classification of sign language image frames before model training (Sneha Prabhu, 2022). Another study introduced a speech-to-sign language system that leverages web scraping techniques to retrieve sign representations from an external repository, specifically the Indian Sign Language Portal. The system completed the conversion task in 28.94 seconds, demonstrating the feasibility of real-time applications using external linguistic resource (Ezhumalai P, 2021).Gesture recognition using CNNs in real time has also gained considerable attention. A system capable of recognizing sign gestures for alphabet letters A–Z achieved an accuracy of 98%. It was further enhanced by integrating predefined American Sign Language (ASL) images and videos for translating audio messages into sign language. This approach highlights the potential of combining gesture classification and media translation for more intuitive communication interfaces. (Pallavi Chaudhari, 2022)

Further advancements have resulted in the development of bidirectional systems that facilitated both Sign Language to Text and Text to Sign Language conversion. A notable implementation involved creating a dataset by capturing gestures using OpenCV with a laptop or webcam and subsequently animating the corresponding gestures through Blender 3D. The system applied TensorFlow and achieved a 90% accuracy rate in text prediction. Feature extraction was enhanced through posture-guided pooling in combination with 3D CNNs, while translation from text to gesture was implemented using NLTK, and audio-to-text conversion utilized the JavaScript Web Speech API (Tanmay Petkar, 2022). These efforts reflect a growing interest in combining machine learning frameworks with real-time multimedia interfaces for gesture communication systems.

A cross-disciplinary approach integrating artificial intelligence techniques with sign language linguistics has been explored to improve recognition systems, emphasizing the importance of addressing the challenges faced by the deaf and hard-of-hearing community in everyday communication (Ronghui Li, 2022). Studies focusing on real-time sign recognition have proposed novel models capable of identifying both alphabets and a subset of characters, achieving an accuracy of 90.78%. Despite some limitations, such frameworks provide reliable two-way communication and serve as foundational efforts for future enhancement in sign gesture coverage and accuracy (Deep Kothadiya, 2022). Further work in this domain employed Natural Language Processing (NLP) techniques along with Google APIs to enable English-to-Indian Sign Language translation. The system achieved 77% accuracy using 3D avatar animation, but its limited vocabulary restricts broader use, highlighting the need for larger lexicons and improved real-time processing (Katta Vaishnavi, 2024).

Recent advancements in sign language recognition (SLR) have increasingly centred on the integration of computer vision, natural language processing, and deep learning methodologies to enhance recognition accuracy, efficiency, and scalability. Research has shown a consistent evolution from traditional handcrafted feature extraction to automated learning through deep neural networks such as CNNs, LSTMs, and Transformers. These architectures enable efficient spatial-temporal representation learning, reducing dependency on manual feature engineering while improving cross-lingual generalization. Moreover, multimodal fusion approaches that combine visual, skeletal, and motion data have significantly improved the robustness and adaptability of recognition systems in real-world environments, contributing to more inclusive communication technologies for the deaf and hard-of-hearing community. (Rupesh Kumar, 2023)

Comprehensive analyses of deep learning techniques for SLR indicate that CNNs remain highly effective for static sign recognition, while LSTM and 3D CNN models excel in capturing temporal dependencies within continuous sign sequences. The adoption of attention mechanisms, transfer learning, and Transformer-based architectures has further enhanced model generalization and translation accuracy across heterogeneous datasets. Frameworks such as Open Pose and Media pipe have become essential in improving spatial

landmark extraction, enabling models to achieve higher recognition stability and performance even under challenging lighting and background conditions. These developments collectively mark a significant shift toward fully automated, data-driven recognition pipelines. (Yulia Kumar, 2024)

The integration of Media pipe with convolutional architectures has proven especially successful for real-time gesture recognition. By extracting twenty-one precise hand landmarks per frame and processing them through CNN-based classifiers, recent systems have achieved remarkably high accuracy, stability, and speed. Through optimized pre-processing, normalization, and data augmentation, such approaches ensure consistent results across diverse users and environments. The combination of lightweight landmark detection and deep spatial learning offers a cost-effective and scalable solution that eliminates the need for additional sensors or wearable devices, making real-time recognition feasible for practical applications in communication and education. (Shahad Thamear Abd Al-Latief, 2024)

Recent transformer-based architectures have emerged as a powerful alternative to conventional deep learning models for sign language recognition. Variants such as Swin and Deformable Attention Transformers leverage hierarchical attention mechanisms and spatial hierarchies to capture fine-grained visual and temporal dependencies across different sign languages. These models have demonstrated exceptional accuracy and generalization on large, diverse datasets while minimizing bias through augmented training and inclusive data collection. Beyond recognition, their adaptability supports multimodal fusion and integration with large language models, enabling future systems capable of multilingual, real-time, and bidirectional sign-to-text and text-to-sign translation (Elvin Lalsiembul Hmar, 2025).

This research builds upon these advancements by integrating Natural Language Processing using Porter Stemming for text simplification, Word2Vec embedding for semantic representation, and a Random Forest Classifier for robust gesture prediction. The resulting output is translated into sign animations using Blender, ensuring accurate and visually interpretable gestures. This integrated pipeline aims to foster inclusivity by bridging communication gaps in healthcare, education, and workplace environments, thus promoting a more accessible future for the hearing-impaired community.

3. Methodology

3.1 Dataset Generation

Dataset generation was the first process undertaken in the development of the English to Sign Language translation system. To ensure accuracy and adherence, the team first learned American Sign Language (ASL) signs from authentic online sources such as Upskill Tutor and ASL Love. An animated character was developed using Blender3D software in which each finger's joints, elbow's joint was manually programmed according to gesture learned from the online sources. Thus, for each alphabet, number and words, the blender's animated character was programmed and rendered to generate accurate ASL gesture.

3.2 Data pre-processing

All of these video clips were encoded at 30 frames per second (FPS) and stored in MP4 file format with an average length of one to three seconds. To expand linguistic coverage, synonyms were mapped to the corresponding gesture videos and incorporated to the dataset. Using this large and ethically obtained dataset as the basis for training the Random Forest classification model, each word was tokenized and transformed into a 100-dimensional numerical vector using Word2Vec. To handle words such as 'happily' and 'happiness', which share the same sign language gesture as 'happy', then Porter Stemming algorithm was implemented.

3.3 Model training

During model training, a Random Forest classifier was used to categorize words into their corresponding sign language gestures. The vectors from Word2Vec served as features (X), while the associated gesture labels acted as target variables (y). Feature extraction involved mapping each word to its corresponding vector representation. The dataset was then split into training and testing sets in an 80:20 ratio where training set contain 100 samples and testing set contain 26 samples respectively. A Random Forest Classifier was initialized with 1000 decision trees and a fixed random state for reproducibility. The model was trained on the feature vectors using the training data. This ensemble approach, leveraging multiple decision trees, helped improve prediction accuracy and minimize overfitting. The result was a robust model capable of accurately predicting the correct sign gesture for a given input word.

Table 1. Dataset Generation

A	After	Again	Against	All	Alone	Also	And	Ask	At
B	Bye	Beautiful	Before	Best	Better	Busy	But	Be	C
Can	Cannot	Change	D	Day	Distance	Do Not	E	F	Faith
From	G	Glitter	Go	God	Gold	Great	H	Hello	Help
Here	His	Home	How	I	J	K	Keep	L	Laugh
Learn	M	Me	More	My	N	Next	Now	O	Of
On	Our	Out	P	Q	R	Right	S	Sad	Safe
See	Self	So	Sound	Stay	Study	T	Talk	Television	Thank You
That	They	This	Those	Time	To	Type	U	Us	V
Walk	Want	Without	Words	Work	World	Wrong	X	Y	You
Yourself	Z	0	1	2	3	4	5	6	7
8	9								

3.4 Recognition

The extracted features from the Porter-Stemming algorithm and test words from the user are converted into tokens using Natural Language Tool Kit (NLTK). Part-of-speech (POS) tags are generated from the tokens and used to retrieve the corresponding sign gestures from the database. The system sequentially reads each sig gesture and renders the corresponding animation until the complete sentence is interpreted.

3.5 Output as Video

The recognized text or letters are displayed sequentially with the corresponding sign language gesture. For example, when the user input ‘Hello’, the system identifies the word and presents tits corresponding sign language gesture. Our system can only process words that are present in our database. If a user inputs a word that is not in our database, our system will generate the sign language gestures for the word by spelling it out letter by letter. For example, if the user types ‘Metabolism’, the model will translate and sequentially display the sign language for each letter: ‘M’, ‘E’, ‘T’, ‘A’, ‘B’, ‘O’, ‘L’, ‘I’, ‘S’, ‘M’.

3.6 Overview of the system

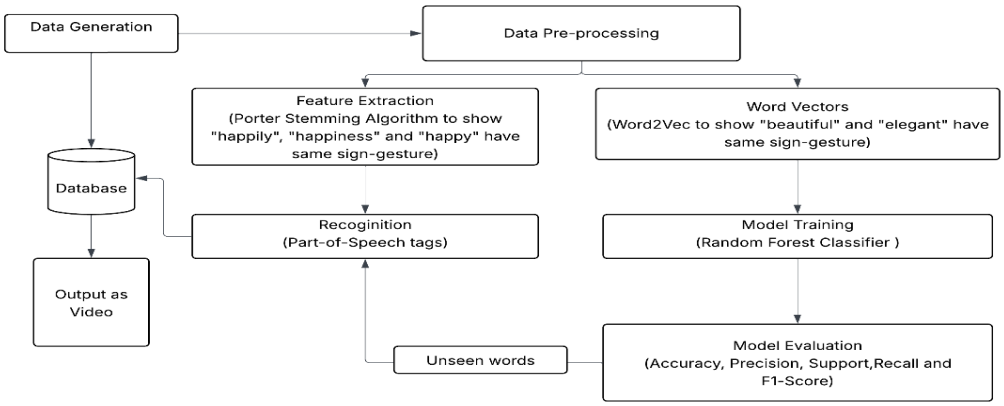


Figure 1. Overview of System

3.7 Performance Evaluation Metrics

The following metrics were considered to evaluate the machine learning model and to analyse its performance.

3.7.1 Accuracy

Accuracy represents the proportion of correctly classified cases, the cases out of the total number of instances in the dataset.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (Equation\ 1)$$

Where:

TP (True Positive) is the number of instances correctly predicted as positive.

FP (False Positive) is the number of instances incorrectly predicted as positive.

TN (True Negative) is the number of instances correctly predicted as negative.

FN (False Negative) is the number of instances incorrectly predicted as negative.

3.7.2 Recall

Recall measures the proportion of true positive cases that the model correctly identifies.

$$\text{Recall} = \frac{TP}{\text{TP} + \text{FN}} \qquad (\text{Equation 2})$$

3.7.3 F1-Score:

The F1 score is a performance metric commonly used in binary classification tasks, which considers both Precision and recall providing a balanced measure of a model's performance.

$$\text{F1} - \text{Score} = \frac{2TP}{\text{2TP+FP+FN}} \qquad (\text{Equation 3})$$

3.7.4 Precision

Precision is a performance metric used in binary classification tasks that measures the proportion of correctly predicted cases out of all cases predicted as positive by the model.

$$\text{precision } = \frac{TP}{\text{TP} + \text{FP}} \qquad (\text{Equaton 4})$$

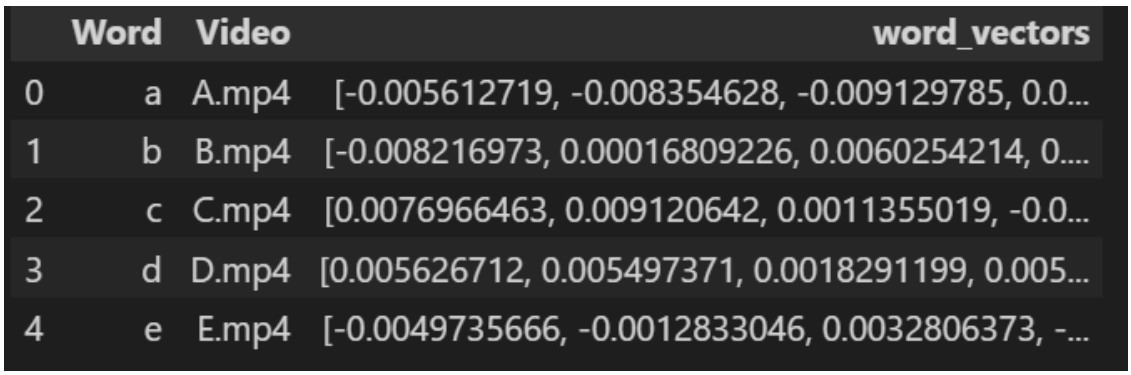
4. Result and Discussion

4.1 Data Generation

A total of 126 sign language gestures were created, comprising numbers 0-9, 26 alphabets, and 90 words. Each sign language gesture video was recorded at a rate of 30 frames per second (fps), meaning that for every second of video, we had 30 individual frames. To expand diversity, we also identified synonyms for 16 words that map to the same ASL gesture.

4.2 Data Pre-processing

After 126 sign language gestures were created, their labels were passed to Word2Vec to generate corresponding word embedding of gesture videos. Thus, 126 such word vectors were made for each label.



	Word	Video	word_vectors
0	a	A.mp4	[-0.005612719, -0.008354628, -0.009129785, 0.0...
1	b	B.mp4	[-0.008216973, 0.00016809226, 0.0060254214, 0....
2	c	C.mp4	[0.0076966463, 0.009120642, 0.0011355019, -0.0...
3	d	D.mp4	[0.005626712, 0.005497371, 0.0018291199, 0.005...
4	e	E.mp4	[-0.0049735666, -0.0012833046, 0.0032806373, -...

Figure 2. Word Vectors

4.3 Mapping of synonyms

Synonyms were managed through a structured semantic mapping framework to ensure that multiple words conveying the same meaning corresponded to a single standardized ASL gesture. During dataset development, we identified that several English words share equivalent sign representations in ASL (for instance, “happy,” “joyful,” and “glad” use an identical gesture). To preserve linguistic consistency, each synonym group was manually verified and assigned to one canonical gesture entry in the dataset. In total, 16 core words within dataset were identified to have synonymous variations. Thus, by applying word embedding (Word2Vec), the system captured contextual relationships among synonyms, allowing closer positioning in vector space for words representing similar meanings. This semantic proximity further reinforced correct class association during Random Forest classification.

4.4 Model’s Performance Evaluation

The model achieved a 94.51% accuracy level, which means that it correctly predicted the outcome in the vast majority of cases. Precision, which is the degree of correctness of optimistic predictions, is optimal (1.0) in most courses, i.e., for all such classes, all of its predictions were correct ones. The word ‘again,’ however, exhibited an exceptionally low precision of 0.18, indicating that that only 18% of its predictions were correct, while the remaining 82% were incorrect.

Table 2. Model Performance Evaluation

Class	Precision	Recall	F1-score	Support
1.mp4	1.0	1.0	1.0	1.0
2.mp4	1.0	1.0	1.0	1.0
3.mp4	1.0	1.0	1.0	2
A.mp4	1.0	1.0	1.0	1
After.mp4	1.0	0.67	0.8	6
Again.mp4	0.18	1.0	0.31	2
B.mp4	1.0	1.0	1.0	2
Beautiful.mp4	1.0	1.0	1.0	3
Before.mp4	1.0	1.0	1.0	3

Recall, which measures the proportion of correctly classified instances, is generally high; however, for the word ‘after’, recall is 0.67, indicating that 33% of its actual instances were not correctly classified. The F1-score, a harmonic mean between Precision and recall, measures moderate performance for ‘after’ (0.80) and low for ‘again’ (0.31), which implies a steep trade-off due to low precision. The support is the actual occurrence of each class in the data set; for instance, ‘after’ has a support of 6, indicating that five of its synonyms appeared in the test set.

Thus, Random Forest Classifier was chosen for its robustness, interpretability, and suitability for small yet diverse datasets. Using Word2Vec-based numerical embeddings as input features, it effectively captured

complex relationships within the data. The ensemble of decision trees ensured stable predictions and strong generalization, even with a moderate dataset size. Compared to deep learning architectures that demand larger datasets and higher computational resources, the Random Forest offered an optimal balance between performance and efficiency, achieving 94.51% accuracy and reliable classification across most gesture classes.

5. Conclusion and Future Enhancements

The deployment of Sign Language Translation System, incorporating the Porter Stemming Algorithm and Random Forest Classifier, successfully integrated text and voice input processing to generate equivalent sign language translations with an accuracy of 94.51%. This represents a significant advancement in bridging communication gaps for deaf and hard-of-hearing individuals. For future improvement of the system, the models can be trained on large and diverse datasets comprising a wide variety of words, phrases, and gestures of sign language. Such enhancement would make the system more scalable, accurate, and performant, making it a more suitable tool for communication in real-world settings.

References

WHO, 2025. [Online] Available at: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss [Accessed 14 Aug 2025].

Danielle Bragg, O. K. M. B. L. B. P. B. A. B. N. C. M. H. H. K. T. V. C. V., 2019. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. ASSETS '19: Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility, pp. 16-31. Necati Cihan Camgoz, S. H. R. B., 2018.

Neural Sign Language Translation. Jie Huang, W. Z. Q. Z. H. L. W. L., 2018. Video-based Sign Language Recognition without Temporal Segmentation. Brian Dickinson, M. G. W. H., 2015.

Dimensionality Reduction of Distributed Vector Word Representations and Emoticon Stemming for Sentiment Analysis. Journal of Data Analysis and Information Processing BREIMAN, L., 2001.

Random Forests. Blender, 2023. [Online] Available at: https://www.blender.org/features/modeling/ [Accessed 14 Aug 2025]. Django, 2023. [Online] Available at: https://www.w3schools.com/django/django_intro.php [Accessed 14 Aug 2025].

Sneha Prabhu, S. S. S. P. S. V. S. J. N., 2022. SIGN LANGUAGE RECOGNITION USING MACHINE LEARNING. International Research Journal of Engineering and Technology (IRJET).

Ezhumalai P, R. K. M. A. S., V. V., Y. A., 2021. Speech To Sign Language Translator For Hearing Impaired. Turkish Journal of Computer and Mathematics Education. Pallavi Chaudhari, P. P. G. M., 2022. Sign Language Detection System. international Journal of Engineering Technology and Management Sciences.

Tanmay Petkar, T. P. A. W. V. C. V. U. D. H., 2022. Real Time Sign Language Recognition System for Hearing and Speech Impaired People. IJRASET. Ronghui Li, L. M., 2022.

Sign language recognition and translation network based on multi-view data. Deep Kothadiya, C. B. ,. S. P.-B. G.-G., M. C., 2022.

Deep sign: Sign Language Detection and Recognition Using Deep Learning. Katta Vaishnavi, M. F. A. B. V. R., 2024. A Natural Language Processing Approach to the Translation of Speech into Indian Sign Language. Rupesh Kumar, A. B. A. S., 2023.

Media pipe and CNNs for Real-Time ASL Gesture Recognition. Yulia Kumar, K. H., -C. L. W. J. L. M. D., 2024. Applying Swin Architecture to Diverse Sign Language Datasets. Shahad Thamar Abd Al-Latief, S. Y. A. A. a. S. K., 2024.

Deep Learning for Sign Language Recognition: A Comparative Review. Elvin Lalsiembul Hmar, B. G. N. R. V., 2025. A Comprehensive Survey on Sign Language Recognition: Advances, Techniques and Applications.