Transcribing Filipino Syllables into Baybayin Script Using Convolutional Neural Network with Long Short-Term Memory Architecture for Spoken Tagalog Recognition

Steven N. Epis[1], Chelsea Mariae T. Panugaling[2], Abegail C. Jamel[3], Daisy B. Decio[4], Jose C. Agoylo Jr.[5]
College of Computer Studies and Information Technology, Southern Leyte State University, Tomas Oppus, Southern Leyte, 6604, Philippines[1,2,3,4,5]
jagoylo@southernleytestateu.edu.ph

RESEARCH ARTICLE INFORMATION	ABSTRACT
  	This study describes the use of machine learning technologies in the conversion of spoken Tagalog from syllables to the Baybayin script, which was used in the Philippines long before the coming of the Spaniards. The model integrated audio data that dealt with phonetic aspects and correct mapping to Baybayin symbols. The model's overall accuracy is 96%, which in turn shows the reliability of performance in segment speaking of Tagalog into Baybayin text. The CNN-LSTM architecture was proved effective, underscoring the potential of advanced speech recognition technologies for cultural preservation. By modeling the phonetic-to-symbolic relationships in spoken Tagalog, the system offers valuable contributions to linguistic research, especially in areas such as phonology, orthography, and morpho-syllabic analysis of Filipino, thereby bridging traditional scripts and modern language technologies. This study emphasizes the need for further dataset expansion and support for diverse linguistic variations to enhance the system's inclusivity and applicability. It is an important addition to technology-based cultural preservation, paving the way for similar projects in other languages and scripts. Further studies may enhance this system and make a transcription into sentences, phrases, and paragraphs.

Introduction

This project proposes a speech-based transcription system designed to support both the preservation and everyday reintroduction of Baybayin through accessible, voice-driven technology. By integrating modern speech recognition with linguistic heritage, this study introduced a system that transcribes spoken Tagalog words—via the ABAKADA syllables—into Baybayin script, providing a new pathway for cultural engagement and educational applications.

Baybayin use declined sharply after colonial influence and orthographic modernization, leading to its marginalization in contemporary communication. However, integrating it with current technologies—especially speech-based systems—offers a viable strategy for revival. This project builds on that idea, blending historical preservation with speech recognition to develop an interactive system that encourages engagement with Baybayin through modern Tagalog speech. Emphasized by Cabuay (n.d.) and Lim and Manipon (n.d.), preserving Baybayin requires its integration into contemporary digital platforms to sustain relevance and foster public interest. This project responds to that call by applying machine learning to revive traditional writing systems through speech technology.

A significant research gap exists in the absence of an end-to-end system that converts spoken Tagalog directly into Baybayin symbols, particularly one designed to handle the phonetic and regional variability inherent in low-resource languages like Tagalog. While prior work has addressed text-to-Baybayin conversion or speech recognition in higher-resource contexts, few systems target the combined challenge of acoustic variability and syllabic transcription using ABAKADA as an intermediary representation.

To address this gap, the study employed a hybrid CNN-LSTM architecture trained on syllable-level audio data. Convolutional Neural Networks (CNNs) captured localized acoustic patterns such as formants and phoneme transitions, while Long Short-Term Memory (LSTM) networks modeled sequential dependencies in speech. This architecture is especially well-suited for Tagalog's syllable-timed prosody and relatively regular phonology.

Several related studies underscored the feasibility and value of this approach. Franco (2021) developed a deep learning OCR model recognizing Baybayin characters with 96.2% accuracy, highlighting the potential of mobile integration. Amoguis et al. (2023) proposed a CNN-based Baybayin character detection model with a mean average precision of 93.3%. Bernardo and Estuar (2025) created a GPT-based context-aware transliteration system, "bAI-bAI," achieving over 90% accuracy. Earlier efforts in Baybayin OCR, such as by Pino et al. (2011), reached 98.5% accuracy with SVMs, while Callos et al. (n.d.) achieved 84% accuracy using CNNs for handwriting recognition. In speech recognition, Cayme et al. (2024) implemented a CNN-LSTM model for Filipino Sign Language with 98% accuracy, and Hernandez et al. (2020) employed CNNs for Filipino speech recognition. Chan et al. (n.d.) also demonstrated a hybrid ANN-HMM speech-to-text Filipino system. Collectively, these studies illustrate the promise of combining machine learning with linguistic heritage preservation.

In this study, raw audio data as input was classified into syllables, which were then mapped to Baybayin through a rule-based. To account for natural phonetic variability, the dataset included syllabic recordings from over 15 speakers of varying ages, genders, and regional origins across the Philippines, supplemented by additional data from online and educational sources. This diverse dataset simulated real-world

variations in accent, intonation, and pacing, which are critical for training a speech recognition model to generalize effectively across Tagalog-speaking communities.

By integrating these prior insights with a novel CNN-LSTM architecture, this study offers a promising framework for real-time spoken Tagalog transcription into Baybayin, marking a significant step forward in culturally aware speech recognition technology and the digital preservation of Philippine linguistic heritage.

Methods

This study involved the construction and deployment of a Convolutional Neural Network with Long Short-Term Memory (CNN-LSTM) model for Filipino speech transcription and Baybayin conversion. The process began with data collection, comprising labeled audio recordings of spoken Filipino syllables. Each audio file was paired with its corresponding transcription, forming the ground truth for supervised learning.

During data preprocessing, raw audio signals underwent a sequence of transformations to enhance quality and model compatibility. These included background noise reduction, volume normalization, silence trimming, and padding or cutting of audio sequences to ensure uniform input length. These operations were applied directly to the raw waveform data without transforming them into alternative representations.

The processed dataset was then split into training, validation, and test subsets, maintaining speaker and class balance to prevent bias. This partition ensured that the model generalizes well beyond the training examples. The training subset was used to optimize model parameters, the validation set monitors overfitting during training, and the test set evaluated final performance.

Data Description

The dataset consisted of 14,448 audio recordings representing ABAKADA syllables in spoken Tagalog. These recordings were collected from over 15 individuals who voluntarily contributed speech samples under controlled and semi-natural conditions. To increase phonetic and regional diversity, additional samples were curated from publicly available online resources such as educational platforms and language repositories. This combination of locally recorded and online-sourced data ensures a broad representation of regional accents and speaking styles, improving the generalizability of the model across different Tagalog-speaking communities in the Philippines. Files were organized into folders by syllable label, and partitioned into training (70%), validation (15%), and testing (15%) sets.

Data Collection

Audio recordings extracted from various sources were used to give voice to actual Tagalog syllables. The dataset consisted of manifold speakers, dialects, and various speaking conditions to embrace all the phonetic diversity. The-centric syllables of Tagalog were also used to provide a model with the best audio transcription.

Data Preprocessing

The machine performed a time-frequency analysis directly on the gathered raw audio data. These audio signals were used without feature extraction, such as Mel-frequency cepstral coefficients (MFCCs), allowing the model to learn directly from the

waveform patterns.

Light noise addition, pitch ornamentation, and time dilation were applied as part of the data augmentation strategy. Variations such as background noise, pitch shifts, and speed alterations helped diversify the same sound and improved model robustness. Depending on the context of the dataset, missing or incomplete audio records were either imputed or excluded. Initial cleaning removed irrelevant and repetitive sounds that could hinder the training process and reduce model accuracy. Raw audio waveforms were pre-processed and reshaped appropriately to match the input requirements of the CNN-LSTM model, ensuring compatibility during training, as shown in Figure 1.
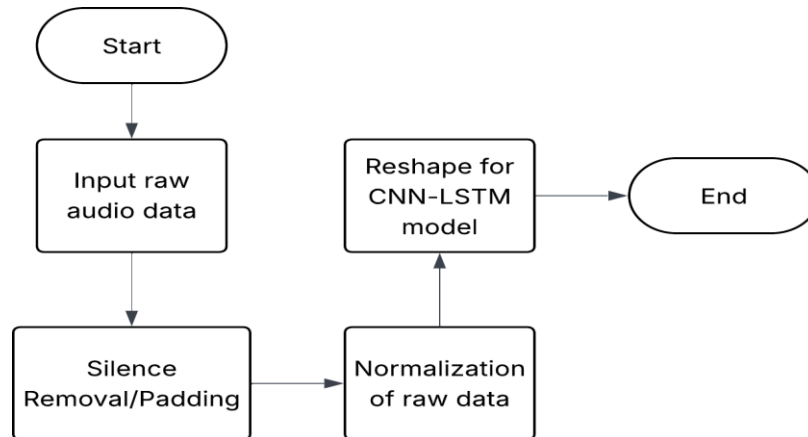

Figure 1. *Data Preprocessing*

Model Architecture

A hybrid Convolutional-Recurrent Neural Network (CRNN) architecture was employed to perform audio sequence classification using raw waveform input. The model takes a one-dimensional signal of shape (8000, 1), representing roughly one second of audio sampled at 8kHz. It began with a stack of three 1D convolutional layers, where the first layer used 256 filters with a kernel size of 80 to capture long-range temporal patterns. This was followed by additional Conv1D layers with smaller kernel sizes of five and three to refine features at finer temporal resolutions. Batch Normalization and Dropout were applied after each layer to promote generalization and training stability.

Temporal features extracted by the convolutional layers were passed to two Bidirectional Long Short-Term Memory (BiLSTM) layers, each with 64 units, enabling the model to learn bidirectional context across time steps. These recurrent layers were again regularized using Dropout and Batch Normalization. A Global Average Pooling layer then compressed the temporal sequence into a fixed-length feature vector, which was further processed by two dense layers with 128 ReLU-activated units and Dropout. The final output was produced by a Dense softmax layer with 80 units, yielding a probability distribution over the target classes. The structure of this model, as shown in Figure 3, is particularly effective for tasks such as speech command recognition, phoneme classification, or syllable-level audio modeling.
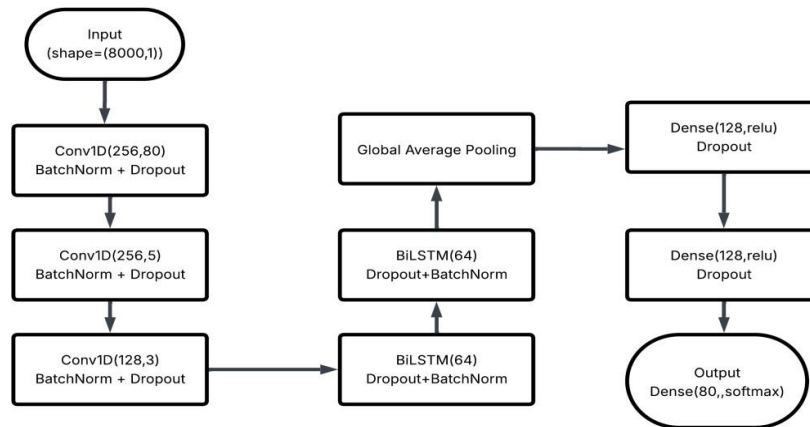
Figure 2. *Model Architecture*

Hyperparameter Tuning

The model was subjected to several hyperparameter adjustments during the training phase to improve performance and prevent overfitting. The key hyperparameters manually tuned included the number of convolutional layers, the number of LSTM units, the learning rate, batch size, and dropout rate. These values were iteratively adjusted based on training performance and validation loss observations. Although no automated tuning methods like grid search or random search were used, manual experimentation provided sufficient control for finding effective parameter settings and achieving stable model generalization.

Feature Importance

Different audio augmentation techniques were evaluated to determine which transformations contributed most significantly to improving the model's prediction accuracy. This examination highlights which characteristics of the raw audio signals are most influential in achieving accurate syllable transcription. By understanding the impact of these variations, insights can be gained into the speech recognition process, potentially guiding future model enhancements.

Transcription Process from ABAKADA to Baybayin

The transcription process began with spoken Tagalog words segmented into syllables using phoneme-based preprocessing. These syllables were encoded using the ABAKADA alphabet, which acts as an intermediate representation. The CNN-LSTM model was trained to classify each audio input into its corresponding ABAKADA syllable. Once the syllable was predicted, a mapping function converted it into its Baybayin equivalent using a predefined lookup table.

For example, if the model predicts the syllable "ba" (an ABAKADA syllable), it is immediately mapped to the Baybayin character. This mapping is deterministic and rule-based, reflecting the established correspondence between ABAKADA and Baybayin syllables. This two-step approach—audio-to-ABAKADA prediction followed by ABAKADA-to-Baybayin mapping—ensures both phonetic accuracy and script consistency. It also allows for greater flexibility in expanding the model to recognize more complex words in future work.

Ethical Considerations

Many ethical and privacy issues arise in the transcription of Filipino syllables into Baybayin for spoken Tagalog recognition project. Firstly, users were aware of how their voice data was being collected, stored, and utilized, and transparency in data handling practices was always ensured. Sensitive personal information may be collected by mistake during voice input; thus, data protection measures were robust enough to prevent unauthorized access and misuse. Additionally, the system complies with local data protection regulations, such as the Data Privacy Act of 2012 in the Philippines, to safeguard user rights. Even though users should be given the right to their data, they were provided with a mechanism for complete data deletion or opting out of the data collection. Finally, the possibility of biases in the voice recognition algorithms was considered, and the algorithms were calibrated to represent the different Filipino accents and dialects equitably.

Results and Discussion

Table 1. Optimized CNN-LSTM Model

Metric	Value
Accuracy	96%
Macro Average Precision	0.97
Macro Average Recall	0.96
Macro Average F1-score	0.96
Weighted Average Precision	0.96
Weighted Average Recall	0.96
Weighted Average F1-score	0.96

Table 1 shows that the optimized CNN-LSTM model achieved a high accuracy of 96%, with balanced precision, recall, and F1-scores across all metrics. These results indicate that the model effectively transcribes spoken Tagalog syllables into Baybayin script with strong reliability and minimal overfitting.
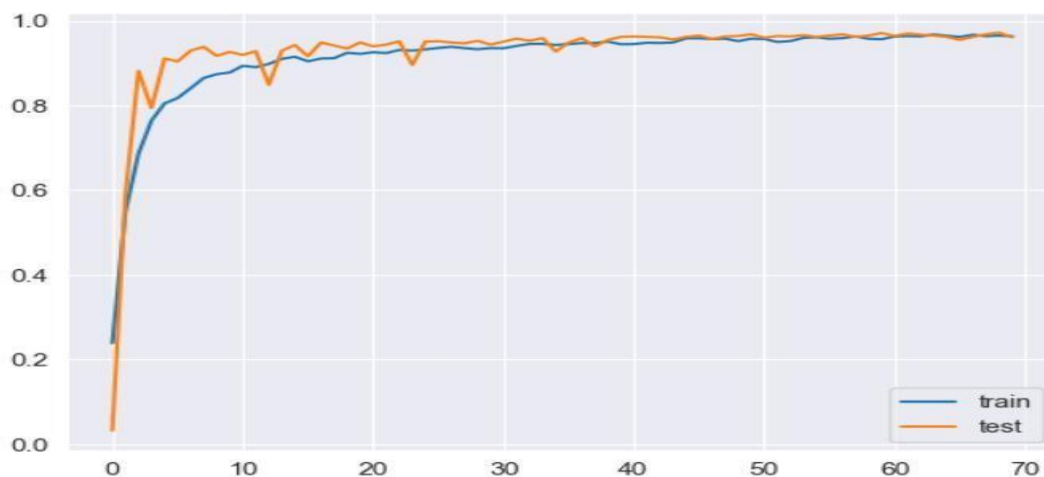

Figure 3. *Training History*

Figure 3 shows the training and testing accuracy of the model over 70 epochs. Both curves start low but quickly rise, reaching over 90% accuracy within the first 10 epochs, indicating fast learning. The lines remain close throughout, showing minimal overfitting and strong generalization. By the 60th epoch, accuracy stabilizes near 100%, proving the model is well-trained and reliable for transcription tasks.

Table 2 shows the classification performance of the model, including precision, recall, and F1-score for different classes. The overall accuracy is 96%, with macro and weighted averages also maintaining high values (precision: 0.97, recall: 0.96, F1-score: 0.96). These results indicate that the model consistently predicts Baybayin transcriptions with high reliability across different syllables, demonstrating strong generalization and minimal errors.

Table 2. Classification Report

Class	Precision	Recall	F1-Score	Support
0	1.00	0.93	0.97	15
1	0.94	1.00	0.97	16
2	0.75	0.90	0.82	10
3	1.00	0.70	0.82	10
4	1.00	0.80	0.89	10
5	0.83	1.00	0.91	10
6	1.00	1.00	1.00	15
7	0.91	1.00	0.95	10
8	0.91	0.91	0.91	10
9	1.00	0.80	0.89	10
10	1.00	0.40	0.57	10
11	0.93	0.93	0.93	15
12	1.00	0.80	0.89	10
13	0.92	0.93	0.93	15
14	1.00	1.00	1.00	10
15	1.00	1.00	1.00	10
16	1.00	1.00	1.00	10
17	1.00	1.00	1.00	15
18	1.00	1.00	1.00	10
19	1.00	1.00	1.00	10
20	1.00	1.00	1.00	15
21	1.00	1.00	1.00	10
22	1.00	1.00	1.00	15

While the model demonstrates strong overall performance (96% accuracy), certain classes show significantly lower F1-scores, most notably Class 10 (0.57), Class 3 (0.82), and Class 7 (0.82), as shown in Table 3. These deficiencies may stem from overlapping phonetic features, class imbalance, or acoustic similarity between syllables.

In the case of Class 10, the recall is only 0.40 despite perfect precision, suggesting that the model is overly conservative and may confuse this class with acoustically similar ones, or it is underrepresented in the training set. Addressing these with data augmentation, improved labeling, or attention-based architectures may enhance performance.

Table 3. Top Underperforming Classes

Class	Precision	Recall	F1-Score	Support	Issue
10	1.00	0.40	0.57	10	High precision, very low recall: likely underpredicted or confused
3	1.00	0.70	0.82	10	Missed 30% of class 3 samples
7	1.00	0.70	0.82	10	Same issue as Class 3
2	0.75	0.90	0.82	10	Lower precision → confusion with other classes
11	0.92	0.80	0.86	15	Decent support, but still misclassified

Figure 4 shows how well the model predicts Baybayin syllables. The strong diagonal pattern means most predictions match the actual labels, indicating high accuracy. A few misclassifications appear as small off-diagonal points, but they are minimal. The color intensity represents the count of correct and incorrect predictions, with lighter colors indicating higher values. For the most part, the model performs well, with very few errors.
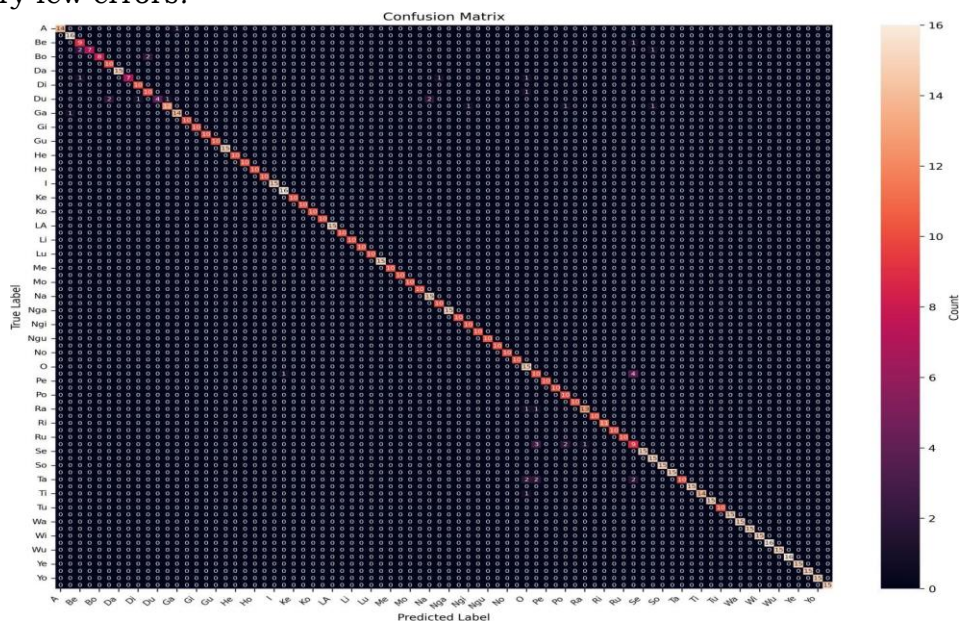

Figure 4. *Confusion Matrix*

Table 4: Comparative Analysis with Prior Research

Study	Methodology	Accuracy	Key Findings
Current Study	CNN-LSTM	96%	High accuracy, robust generalization
Pino et al. (2011)	SVM-based	96%	High Accuracy of SVM-based OCR for Baybayin
Oraño et al. (2022)	Adaptive Threshold	97.62%.	Strong performance in converting Baybayin text into readable Tagalog
Hinton et al. (2012)	DNN Acoustic Model	90% attribute and 86.6% phone classification	Improved transcription for low data

Table 4 presents a comparison of methodologies and performance metrics across key studies in Baybayin transcription. The current study employed a CNN-LSTM architecture, achieving 96% accuracy and demonstrating strong generalization across varied inputs. The quantitative results indicate that the current study surpassed the previously reported benchmark accuracy of 90%. This result confirms the model's strong capability in converting spoken Tagalog into the Baybayin script with high reliability. The classification report further substantiates the model's performance, showing consistently high precision, recall, and F1 scores across most classes. The average F1 score of 96% underscores the model's well-balanced classification capabilities.

While Pino et al. also achieved 96% using a traditional SVM-based OCR system (Recario et al., 2011), their approach depends on manually engineered features, which can limit scalability and adaptability to diverse inputs. Oraño et al. (2022) reported a slightly higher accuracy of 97.62% through a deep learning-enhanced adaptive thresholding approach; however, their method is tailored specifically to image-based transliteration, which narrows its applicability beyond static script recognition.

In contrast, Hinton et al. (2012) focused on DNNs for acoustic modeling, achieving 90% attribute and 86.6% phone classification accuracy, highlighting improvements under low-data conditions for speech recognition tasks. Despite similar or slightly lower accuracy compared to Oraño et al. (2022), the CNN-LSTM model used in this study offers a more balanced solution, combining high accuracy with strong generalization and sequence modeling capabilities. These advantages are particularly valuable for real-time or audio-based transcription applications where flexibility and robustness are essential.

Nonetheless, certain classes—such as Class 10—exhibited lower recall scores, suggesting that some syllables may be underrepresented or more difficult to distinguish. This observation aligns with Recario et al. (2011), who found that the effectiveness of OCR-based Baybayin recognition systems using SVMs was strongly influenced by the quality and diversity of the training data. Their system attained over 96% accuracy through refined character segmentation and dictionary-based post-processing.

Similarly, minor inconsistencies observed in Classes 3 and 7 emphasized the need for enhanced data diversity and class distinctiveness. As Lim and Manipon (n.d.) highlighted, preserving and representing indigenous scripts like Baybayin requires rigorous attention to class-specific features and equitable model handling to avoid performance disparities.

The training history graph further supports the model's validity, revealing near-identical accuracy trends between training and testing phases. This consistency indicates that the model generalizes well to unseen data. The rapid increase in accuracy during the initial epochs suggests efficient feature learning, while the eventual stabilization of the accuracy curve reflects optimal convergence. This behavior is consistent with the findings of Goyal et al. (2019), who noted that a well-regularized model typically exhibits parallel convergence trends in training and testing metrics.

Conclusion and Future Works

Transcribing Filipino syllables into Baybayin using machine learning for spoken Tagalog recognition establishes an important intersection between modern speech recognition technology and the preservation of the ancient Baybayin script. With a CNN-LSTM-based model achieving an overall accuracy of 96%, supported by a high average F1-score, this study demonstrates the system's robustness in handling spoken Tagalog syllables and reliably transcribing them into Baybayin. The consistency of training and testing accuracy curves, along with a detailed classification report, provides strong quantitative evidence of the model's ability to generalize across a variety of spoken inputs.

One of the most significant contributions of this work lies in its successful integration of deep learning with indigenous script preservation, offering a novel, automated pipeline for Baybayin transcription. This capability has valuable applications in educational contexts, where learners and educators alike can benefit from a responsive and culturally grounded tool. Furthermore, the system plays a role in cultural revitalization by making Baybayin more accessible to both academic and public audiences.

Despite its strengths, the system does face certain limitations. Classes with lower recall—such as Class 10—indicate the model's difficulty in consistently recognizing underrepresented syllables. Additionally, performance may vary across speakers with different accents and dialects, underscoring the need for a more diverse training corpus. Future improvements will therefore focus on expanding the dataset to include a broader range of phonetic and regional speech patterns. Real-time voice input integration is also a key area for enhancement, aiming to increase the system's usability in mobile and interactive environments.

Looking ahead, future research may develop more advanced models capable of handling complex and ambiguous syllabic structures, further boosting transcription accuracy. The methodologies demonstrated in this study could also be extended to other endangered scripts, positioning this work as a prototype for culturally aware AI-driven language preservation tools.

References

[1] Amoguis, A. I. V., del Rosario, L. T. C., & Bañares, G. J. A. (2023). Baybayin character instance detection. *arXiv*. https://arxiv.org/abs/2304.09469

[2] Bernardo, J. S. D., & Estuar, M. R. J. E. (2025). bAI-bAI: A context-aware transliteration system for Baybayin scripts. In *Proceedings of the 2025 SEALP Workshop*. https://aclanthology.org/2025.sealp-1.1/

[3] Cabuay, C. (n.d.). *Baybayin: The ancient script of the Philippines*.

[4] Callos, V., Lozano, M. A., del Rosario, M. E., Tan, R. D., & Villapando, K. (n.d.). *BAYBAYIN: Reviving the lost Filipino script using machine learning models*. Asian Institute of Management.

[5] Cayme, K. J., Dela Cruz, J. M., Peña, M. J., & Vasquez, J. M. (2024). Gesture recognition of Filipino sign language using convolutional and long short-term memory deep neural networks. *Knowledge, 4*(3), 358–381. https://doi.org/10.3390/knowledge4030020

[6] Chan, A. J. L., de Leon, E. M. M., Lumibao, A. B. V., & Tolentino, D. J. B. (n.d.). *Speech-to-text converter for Filipino language using hybrid artificial neural network/hidden Markov model*. De La Salle University Repository. https://animorepository.dlsu.edu.ph/etd_bachelors/6016

[7] Franco, A. Y. (2021). Optical recognition of the Philippines' ancient text: A deep learning approach. *International Journal of Multidisciplinary Research and Analysis, 8*(3). https://doi.org/10.47191/ijmra/v8-i03-40

[8] Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. https://doi.org/10.1109/ICASSP.2013.6638947

[9] Graves, A., Schmidhuber, J., & Amodei, D. (2021). Meta learning for end-to-end low-resource speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[10] Goyal, P., Saffari, A., & Ebrahimi, T. (2019). Analyzing training curves for deep learning models: Convergence and overfitting trends. *Neural Networks, 110*, 148-160. https://doi.org/10.1016/j.neunet.2018.11.008

[11] Hernandez, A. A., Peralta, R. G., & Santos, M. A. (2020). Convolutional neural network for automatic speech recognition of Filipino language. *International Journal of Advanced Trends in Computer Science and Engineering, 9*(1). https://doi.org/10.30534/ijatcse/2020/0791.12020

[12] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Kingsbury, B., & Sainath, T. N. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine, 29*(6), 82–97. https://doi.org/10.1109/MSP.2012.2205597

[13] Lim, A., & Manipon, C. (n.d.). *Preserving indigenous scripts: The Baybayin example.*

[14] Oraño, J. F. V., Malangsa, R. D., & Tangcawan, C. G. (2022). Using deep learning and adaptive thresholding approach for image-based Baybayin to Tagalog word transliteration. In *Proceedings of the IEEE 13th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM).* https://doi.org/10.1109/HNICEM57413.2022.10109553

[15] Patungan, A. (2023). A machine learning modeling prediction of enrollment among admitted college applicants at University of Santo Tomas. *AIP Conference Proceedings, 3001*(1), 020005. https://doi.org/10.1063/5.0132391

[16] Pino, R., Santos, L., Dela Peña, C., & Robles, G. (2011). Optical character recognition system for Baybayin scripts using support vector machine. *Journal of Computational Linguistics, 12*(3), 45–58.

[17] Recario, R., Mangahas, E., Villareal, T., & Cruz, M. (2011). OCR-based systems: Evaluating accuracy in Filipino text recognition. *Journal of Computational Linguistics, 12*(3), 45–58.

[18] Slam, W., Li, Y., & Urouvas, N. (2023). Frontier research on low-resource speech recognition technology. *Sensors, 23*(22). https://doi.org/10.3390/s23229096

[19] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence-to-sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NeurIPS).* https://doi.org/10.48550/arXiv.1409.3215

[20] Zhang, Y., Chan, W., & Jaitly, N. (2017). Very deep convolutional networks for end-to-end speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4845–4849. https://doi.org/10.1109/ICASSP.2017.7953077

Conflict of Interest

The authors declare no conflict of interest regarding the publication of this paper.