engineering proceedings

MDPI

Proceeding Paper

Utilizing Residual Network 50 Convolutional Neural Network Architecture for Enhanced Philippine Regional Language Classification on Jetson Orin Nano †

John Paul T. Cruz [1,*], Aaron B. Abadiano [1], FP O. Sangilan [1], Emmy Grace T. Requillo [2,*] and Roben A. Juanatas [3]

[1] School of Electrical, Electronics, and Computer Engineering, Mapúa University, Manila 1002, Philippines; ababadiano@mymail.mapua.edu.ph (A.B.A.); fosangilan@mymail.mapua.edu.ph (F.O.S.)
[2] College of Engineering and Architecture, Mapúa Malayan Colleges Mindanao, Davao City 8000, Philippines
[3] College of Computing and Information Technologies, National University, Manila 1008, Philippines; rajuanatas@national-u.edu.ph
* Correspondence: jptcruz@mapua.edu.ph (J.P.T.C.); egrequillo@mcm.edu.ph (E.G.T.R.)
† Presented at the 7th Eurasia Conference on IoT, Communication and Engineering 2025 (ECICE 2025), Yunlin, Taiwan, 14–16 November 2025.

Abstract

Visual speech recognition systems encounter significant challenges in multilingual nations such as the Philippines, where numerous regional languages, including Cebuano and Ilocano, feature distinct phonetic-visual characteristics. Deep learning models such as the Lip Reading Network and the Lightweight Crowd Segmentation Network have demonstrated strong performance with 3D Convolutional Neural Networks (CNNs). However, their substantial computational requirements restrict deployment on portable edge devices. We introduce a more efficient alternative that integrates a 2D Residual Network 50 architecture with a Long Short-Term Memory network and Connectionist Temporal Classification for lip-reading classification of Philippine regional languages. The proposed model is deployed on the Jetson Orin Nano, a high-performance edge device optimized for real-time inference through Compute Unified Device Architecture acceleration. Using a dataset of 2000 annotated videos encompassing 10 lexicons each for Cebuano and Ilocano, the model's effectiveness was evaluated. Results achieved a regional language classification accuracy of 90%, with lexicon-level accuracies of 74% for Cebuano and 66% for Ilocano. This work represents a step toward developing accessible and scalable communication aids for deaf communities in linguistically diverse environments, leveraging transfer learning on pretrained models.

Keywords: visual speech recognition; ResNet50; transfer learning; edge computing; Jetson Orin Nano; Philippine regional languages; lip-reading

Academic Editors: Teen-Hang Meen, Chi-Ting Ho and Cheng-Fu Yang

Published: 26 March 2026

1. Introduction

Visual speech recognition, particularly lip-reading, provides an essential communication pathway for individuals with hearing impairments by interpreting visemes as visual correlates of speech. In multilingual contexts such as the Philippines, which comprises over 150 regional languages and dialects [1], communication challenges are particularly pronounced. Practices such as code-switching, diverse tonalities, stress contrasts, and vowel elongation further complicate the linguistic landscape [2]. Among the most widely

spoken regional languages are Cebuano and Ilocano [3], both of which exhibit distinct phonetic-visual characteristics.

Current visual speech recognition systems predominantly focus on widely spoken languages, thereby neglecting regional dialects due to limited training data and the inability to capture language-specific phonetic nuances. This exclusion results in disparities in assistive technologies and accessibility opportunities for local deaf communities. Addressing these challenges through the development of optimized models extends the benefits of AI to underserved populations.

This study aims to contribute to this effort by proposing an efficient, real-time lip-reading system for Philippine regional languages deployable on portable edge devices. Specifically, the objectives are: (1) to develop a Jetson Orin Nano-based device for capturing lip movements; (2) to implement a 2D Residual Network 50 (ResNet50) architecture combined with a Long Short-Term Memory (LSTM) network and a Connectionist Temporal Classification (CTC) decoder for dialect and lexicon classification; and (3) to evaluate the model's accuracy and inferencing performance using a dataset of 2000 annotated videos comprising 10 lexicons each for Cebuano and Ilocano.

2. Literature Review

The advent of deep learning has significantly transformed visual speech recognition, with modern systems employing neural networks to extract complex visual features and temporal dependencies. These advancements are evident across diverse applications in classification, identification, and detection tasks. For instance, deep learning models have been applied to distinguish Arabic speech dialects [4] and to classify natural versus synthetic dyes in textiles using Residual Neural Networks with transfer learning [5]. In agriculture, Convolutional Neural Networks (CNNs) have been used to classify healthy and diseased Abaca leaves [6], while the You Only Look Once (YOLO) algorithm has been employed to detect pests such as whiteflies and fruit flies [7]. Beyond classification, CNNs have facilitated identification tasks, including recognizing medicinal mushrooms [8], assessing fish freshness through gill color analysis [9], and diagnosing Abaca diseases [10]. Deep learning has also extended to non-visual domains, such as urban sound identification using hybrid LSTM-Support Vector Machine (SVM) models with haptic feedback [11]. Detection applications include Mask R-CNN architectures for identifying coffee beans [12] and Philippine coins [13], as well as public safety systems for wireless fire detection [14] and retail innovations in automated checkout systems on edge devices like the NVIDIA Jetson Nano [15].

In lip-reading, Lip Reading Network (LipNet) [16] and the Lightweight Crowd Segmentation Network (LCSNet) [17] have demonstrated notable performance using 3D CNNs. These architectures are particularly effective for spatiotemporal feature extraction, capturing both static and dynamic lip movements across video sequences. Similarly, CTC has been employed for the sub-classification of regional languages. Despite their effectiveness, 3D CNNs are computationally intensive, especially when processing high-resolution video inputs or extended sequences. In contrast, 2D CNNs, which emphasize local spatial information, have achieved significant breakthroughs in image recognition tasks, such as fish freshness classification [9].

To address computational efficiency, we replace integrated 3D CNNs with a 2D ResNet50-based transfer learning approach. ResNet, introduced by Microsoft in 2015 [18], was designed to overcome the vanishing gradient problem in deep networks. ResNet50, a variant with 50 layers, incorporates residual blocks that create shortcut connections, enabling inputs to bypass certain layers and be directly added to subsequent outputs. This mechanism facilitates gradient flow, thereby improving the performance of deeper models

https://doi.org/10.3390/engproc2026134002

without sacrificing accuracy. While ResNet50 is designed for static image recognition, its integration into visual speech recognition allows spatial feature extraction from individual video frames. These features can then be processed by a bidirectional LSTM to capture the temporal dynamics of lip movements. By cascading spatial and temporal models, the system leverages the strengths of both architectures to achieve efficient and accurate lip-reading.

3. Methodology

3.1. System Workflow

Figure 1 illustrates the system workflow of this study, highlighting the input, processing stages, and output of the proposed device. The input consists of a three-second video recorded at 24 frames per second using a Logictech Brio 100 1080p camera (manufactured by Logitech Europe, Lausanne, Switzerland) connected to and powered by the Jetson Orin Nano. The video undergoes preprocessing, wherein the Dlib68 facial landmark detector is employed to isolate the mouth region of the speaker. To ensure compatibility with the ResNet50 architecture pretrained on ImageNet, the video frames are converted from monochrome to RGB format.

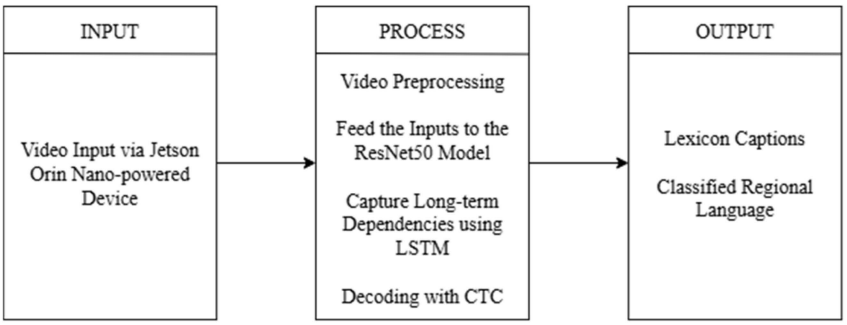

Figure 1. System workflow.

The preprocessed frames are then passed into the model architecture, which utilizes ResNet50 layers for per-frame feature extraction in place of a three-dimensional convolutional layer. These features are processed through a time-distributed layer to preserve frame-wise information, followed by a recurrent neural network (RNN). In this study, an LSTM network is employed to capture long-term temporal dependencies and mitigate the vanishing gradient problem. The LSTM outputs are subsequently fed into a CTC layer, which performs regional language and lexicon recognition.

3.2. System Component

The system component in the visual speech recognition system is presented in Figure 2. This system incorporates a Jetson Orin Nano as an edge device for the classification of Philippine Regional Languages. The Jetson Orin Nano is connected to a 12 V power supply provided with the Carrier Board Kit of the device. It is also connected to a MLE01284 7-inch capacitive touch LCD screen for Raspberry Pi, manufactured by Makerlab PH (Manila, Philippines), for the user to interact with and to display the output. The microcomputer is connected to a 1080 pixel camera for input, positioned below the LCD screen primarily to capture the mouth. The enclosure has a light source to ensure proper and consistent lighting for the user of the device. This configuration was used both for data gathering and testing to ensure a proper and consistent dataset for testing processes.

https://doi.org/10.3390/engproc2026134002

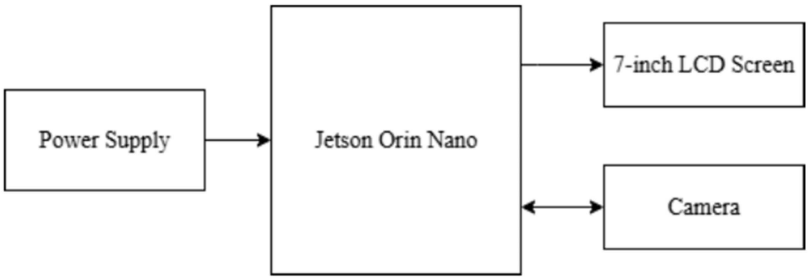

Figure 2. System component.

3.3. Model Architecture

The model architecture illustrates the processes of both learning and inference (Figure 3). The input consists of frames extracted from user-captured videos, which undergo preprocessing before being fed into the ResNet50 layer. This convolutional neural network associates visual features from the frames with corresponding lip movements. The extracted features are then passed to the LSTM module, which provides temporal context by modeling the sequential arrangement of lip movements. This enables the system to identify which movements co-occur to form specific words. Subsequently, a linear transformation is applied to convert the LSTM outputs into an output tensor through a weighted linear operation. Finally, the CTC layer ensures accurate word prediction by preventing redundant character repetitions, which may arise when multiple frames correspond to the same letter.

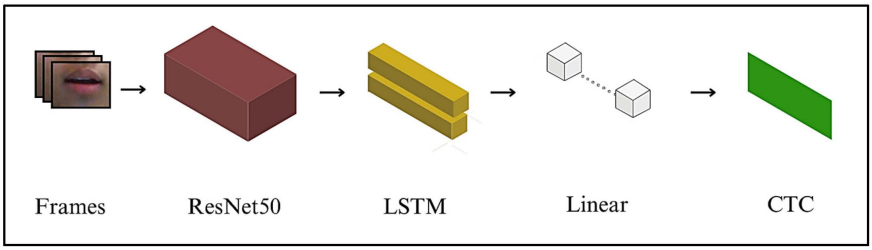

Figure 3. Model Architecture via Transfer Learning.

Figure 4 depicts the complete ResNet50 architecture, which comprises a series of convolutional blocks organized into four stages of residual learning. The initial stage begins with a 7×7 convolutional layer that downsamples the input image while extracting low-level features such as edges and textures. The use of 64 filters establishes the initial channel depth for subsequent layers. In the first stage, residual connections are introduced to maintain spatial dimensions while enabling residual learning. Each block employs a bottleneck design, wherein convolutional layers compress the channels and extract spatial features from the region of interest. The second stage increases feature complexity while reducing input resolution. The third stage captures high-level features, including object shapes and global patterns. The fourth stage prepares the features for classification at the highest level of abstraction. Here, global average pooling reduces each 2048-channel feature map to a single value, rendering the network invariant to input size. Finally, a fully connected layer maps the pooled features to class scores.

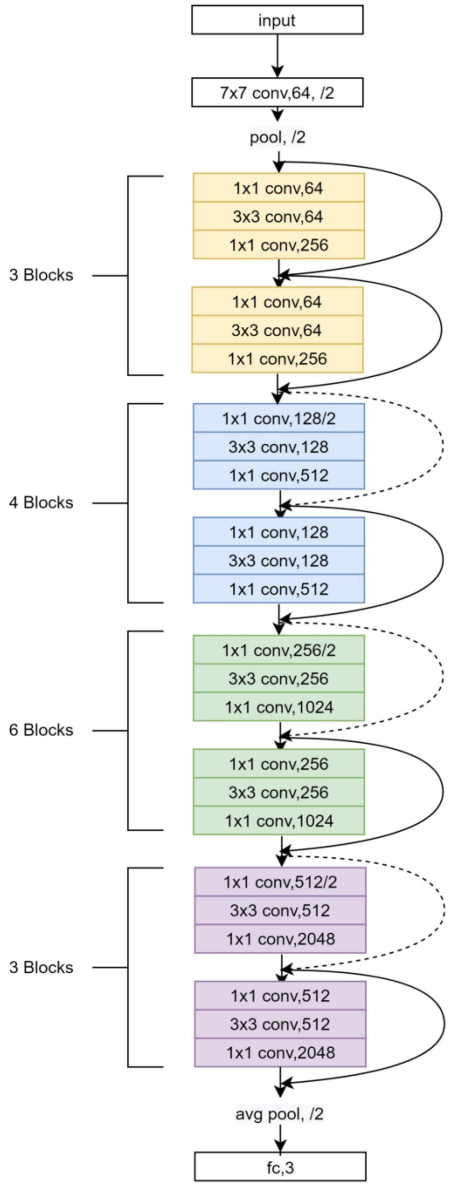

Figure 4. ResNet50 architecture.

3.4. Software

Figure 5 presents the process of the software, with all processing tasks executed on the Jetson Orin Nano. The workflow begins with video capture using the device described in Figure 2. The recorded video is preprocessed within the data_preprocess() module, where it is divided into individual frames and the facial region of interest (ROI) is isolated using Dlib's 68-point landmark detection library. The preprocessed frames are then passed to the self-trained transfer learning model within the model_processing() module, which employs ResNet50 pretrained on ImageNet for feature extraction.

Following feature extraction, the data is processed through a time-distributed layer to reduce dimensionality and enhance feature discrimination. The resulting sequence is subsequently fed into an LSTM layer, which performs spatio-temporal modeling and retains contextual information across consecutive frames. The LSTM outputs are then passed to a CTC layer, which generates the final predictions while mitigating redundancy caused by repeated characters across frame sequences. The program flow diverges depending on whether the user initiates training or inference. In the training phase, the model_training() module defines the model architecture, applies user-specified hyperparameters, and itera-

tively optimizes performance across multiple epochs until satisfactory accuracy is achieved. In the inference phase, the model_inferencing() module applies the trained weights to the defined model in the model_processing() module and performs prediction. The final output consists of the recognized lexicon and the classified regional language.

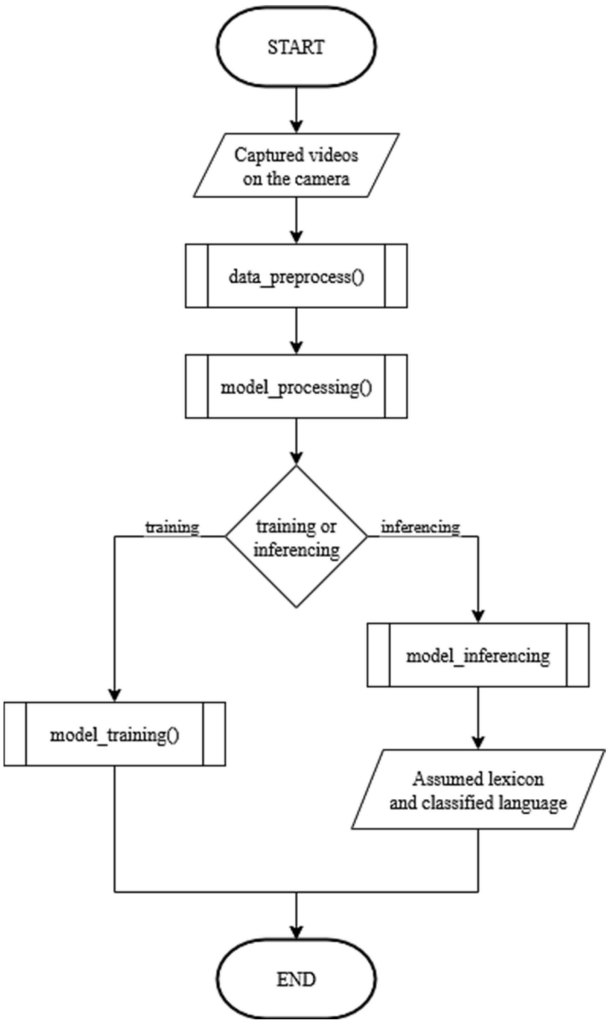

Figure 5. Process of software.

3.5. Experimental Setup

The experimental setup for visual speech classification using transfer learning from ResNet50 on Philippine regional languages is illustrated in Figure 6. The devices are housed within a solid black acrylic enclosure measuring 15 × 11.5 × 11.5 inches to minimize reflections. The process begins by powering on the Jetson Orin Nano and activating the supplementary light source, also enclosed in black acrylic. The software, developed in Python 3.10 using Tkinter 8.6.14 as the graphical user interface (GUI), is then launched to enable testing of the trained lip-reading model for data acquisition. The users position themselves so that the mouth region is visible within a green bounding box displayed on the screen. Once visibility is confirmed, the "Start Inferencing" button on the 7-inch LCD screen is pressed to initiate the process. The model requires approximately three to seven seconds to complete inference, after which the recognized output is displayed on the LCD screen.

(**a**) (**b**)

Figure 6. Experimental Setup using the developed prototype showing (**a**) an off-light source and (**b**) an on-light source.

3.6. Data Gathering and Analysis

A total of 2000 videos were collected, each lasting three seconds and recorded at 24 frames per second. These videos were divided into two regional languages, Cebuano and Ilocano. Each language comprised ten lexicons corresponding to direct translations of the following English phrases: good morning, good afternoon, good evening, goodbye, I am fine, please, take care, you're welcome, what are you doing? and thank you very much.

To evaluate model performance, a 10 × 10 multiclass confusion matrix was employed, corresponding to the ten lexicons for each regional language. This matrix provides a comprehensive representation of classification outcomes by mapping predicted labels against actual labels across all classes. It highlights correctly classified instances (true positives) as well as misclassifications (false positives and false negatives). From this matrix, the overall system accuracy was computed using the standard formula for multiclass accuracy.

$$\% \, Accuracy = \frac{\sum_{n=1}^{10} A_{nn}}{\sum_{\substack{i=1 \\ j=1}}^{10} A_{ij}} \times 100 \quad (1)$$

4. Results and Discussion

The model achieved an accuracy of 96% for Cebuano lexicon recognition, with 48 correct predictions and 2 incorrect predictions. As for Ilocano, the model showed an accuracy of 86% with 42 correct classifications and 8 incorrect classifications. The detailed performance is illustrated in the confusion matrix in Table 1. The regional language classification yielded an average accuracy of 90%.

Table 1. Confusion Matrix for Regional Language Classification.

		Observation	
		Cebuano Lexicons	Ilocano Lexicons
Prediction	Cebuano Lexicons	48	8
	Ilocano Lexicons	2	42

https://doi.org/10.3390/engproc2026134002

For Cebuano, these lexicons directly translate to: C1—Ma-ayong Buntag, C2—Ma-ayong Hapon, C3—Ma-ayong Gabii, C4—Amping, C5—Maayo man ko, C6—Palihug, C7—Mag-amping ka, C8—Wala'y sapayan, C9—Unsa imong ginabuhat? C10—Daghang Salamat!

For the Cebuano lexicons, the model achieved a lexicon-level accuracy of 74.00%, with 37 correct classifications and 13 misclassifications. A key insight from the error analysis relates to the most frequently occurring phrases in the dataset. Greetings beginning with "Maayo"—specifically "Maayong buntag", "Maayong hapon", and "Maayong gabii"—were among the most common phrases. This linguistic pattern directly influenced model performance, as a substantial number of misclassifications occurred among these greetings. The likely cause is that the initial viseme sequence for "Maayo-" is identical across these phrases, making them difficult to distinguish until the final syllables (-tag, -pon, -bii), which are less visually emphasized. This limitation contributed to the errors documented in Table 2. Another source of misclassification was the high frequency of the phrase "Unsa imong buhaton?" As a multi-word expression, it posed challenges for the model when distinguishing it from other longer phrases, further contributing to the overall error rate. For Ilocano, the lexicons used in the dataset were as follows: L1—Naimbag a bigat, L2—Naimbag a malem, L3—Naimbag a rabii, L4—Diyos iti agyaman, L5—Mayat met, agyamanak, L6—Paki, L7—Ag im-imbag ka, L8—Awan ti ania, L9—Ana't ub-ubraem, and L10—Agyamanak un-unay!

Table 2. Confusion matrix for cebuano lexicon accuracy.

		Observation									
		C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Observation	C1	5	0	0	0	2	0	0	0	0	0
	C2	0	5	0	0	1	0	0	0	0	0
	C3	0	0	5	0	0	0	0	0	0	0
	C4	0	0	0	4	0	0	0	0	0	0
	C5	0	0	0	0	2	0	0	0	0	0
	C6	0	0	0	0	0	4	0	0	0	0
	C7	0	0	0	0	0	0	4	0	0	2
	C8	0	0	0	0	0	0	0	3	0	0
	C9	0	0	0	0	0	1	0	2	3	0
	C10	0	0	0	0	0	0	1	0	2	2

In Table 3, the confusion matrix reveals a situation parallel to that observed in Cebuano, with 33 correct classifications and 17 misclassifications, resulting in an overall accuracy of 66.00%. The most frequent Ilocano phrases are dominated by greetings that share a common root: "Naimbag a bigat", "Naimbag a malem", and "Naimbag a rabii". The shared initial visemes of "Naimbag a" introduce substantial visual ambiguity for the lip-reading model, leading to confusion among these distinct greetings. The prevalence of the phrase "Anat ub-ubraem" mirrors the frequency of its Cebuano counterpart. Although the lexicons differ, the articulatory pattern of multi-syllable questions may exhibit similar visual characteristics, potentially contributing to inter-language confusion during regional language classification. This suggests that the model may be learning rhythmic or prosodic similarities rather than focusing exclusively on language-specific phonetic details. In contrast, the frequent occurrence of short, simple words such as "Paki", which possess unique visemes, likely accounts for some of the model's more accurate predictions within the Ilocano dataset.

Table 3. Confusion matrix for ilocano lexicon accuracy.

		Observation									
		L1	L2	L3	L4	L5	L6	L7	L8	L9	L10
Observation	L1	2	2	0	0	0	0	1	0	0	0
	L2	3	3	0	0	0	0	0	0	0	0
	L3	0	0	3	0	0	0	1	0	0	0
	L4	0	0	0	4	0	0	0	0	0	0
	L5	0	0	0	0	2	0	0	0	0	0
	L6	0	0	0	0	0	3	0	0	0	0
	L7	0	0	0	0	0	0	3	0	0	2
	L8	0	0	0	0	0	0	0	5	0	0
	L9	0	0	0	0	0	0	0	2	5	2
	L10	0	0	0	0	0	0	0	0	0	3

5. Conclusions and Recommendations

The system developed demonstrates the effective application of a pre-trained ResNet50 model for the visual-only classification of Philippine regional languages, achieving 90% accuracy in distinguishing Cebuano from Ilocano on a Jetson Orin Nano prototype. While the model performed well at the macro-level classification task, lexicon-level accuracies of 74% for Cebuano and 66% for Ilocano revealed a key challenge: the misclassification of common greetings that share identical initial visemes, such as the "Naimbag a-" roots. This finding highlights the need for more robust sequential analysis to better differentiate visually similar phrases. These limitations can be addressed by exploring hybrid architectures that integrate ResNet50 with Transformer or LSTM layers to capture finer temporal cues. Expanding the training dataset beyond frequently used phrases would help mitigate model bias, while advanced fine-tuning techniques could further enhance classification precision.

Author Contributions: Conceptualization, J.P.T.C., A.B.A.,F.O.S., E.G.T.R., and R.A.J.; methodology, A.B.A. and F.O.S.; software, A.B.A. and F.O.S.; validation, J.P.T.C., E.G.T.R., and R.A.J.; formal analysis, A.B.A. and F.O.S.; investigation, J.P.T.C., A.B.A. and F.O.S.; resources, J.P.T.C., E.G.T.R. and R.A.J.; data curation, A.B.A. and F.O.S.; writing—original draft preparation, J.P.T.C. and F.O.S.; writing—review and editing, J.P.T.C., A.B.A. and F.O.S.; visualization, J.P.T.C., E.G.T.R., and R.A.J.; supervision, J.P.T.C., E.G.T.R. and R.A.J. project administration, J.P.T.C., E.G.T.R. and R.A.J.; funding acquisition, E.G.T.R. and R.A.J. All authors have read and agreed to the published version of the manuscript.

Funding: The presentation was funded by Mapua Malayan Colleges Mindanao through the Office for Research, Development, and Innovation.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Datasets were obtained from selected sites by the researchers. These may be requested by sending an e-mail to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Amorim, R. Code switching in student-student interaction: Functions and reasons! *Linguistica* **2012**, *7*, 177–195.
2. Reid, L.; Rubino, C. Ilocano Dictionary and Grammar: Ilocano-English, English-Ilocano. *Ocean. Linguist.* **2002**, *41*, 238–243. [CrossRef]
3. Wolff, J.U. *A Dictionary of Cebuano Visayan*; Neobooks: Munich, Germany, 1972.

4. Alrehaili, M.; Alasmari, T.; Aoalshutayri, A. Arabic speech dialect classification using deep learning. In Proceedings of the 2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC), Jeddah, Saudi Arabia, 23–25 January 2023; pp. 1–5. [CrossRef]
5. Zamudio, K.J.D.; Caya, M.V.C. Natural or synthetic dye classification in textiles using residual neural network and transfer learning. In Proceedings of the 2022 IEEE 14th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), Boracay Island, Philippines, 1–4 December 2022; pp. 1–5. [CrossRef]
6. Buenconsejo, L.T.; Linsangan, N.B. Classification of healthy and unhealthy abaca leaf using a convolutional neural network (CNN). In Proceedings of the 2021 IEEE 13th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), Manila, Philippines, 28–30 November 2021; pp. 1–5. [CrossRef]
7. Legaspi, K.R.B.; Sison, N.W.S.; Villaverde, J.F. Detection and classification of whiteflies and fruit flies using YOLO. In Proceedings of the 2021 13th International Conference on Computer and Automation Engineering (ICCAE), Melbourne, Australia, 20–22 March 2021; pp. 1–4. [CrossRef]
8. Sutayco, M.J.Y.; Caya, M.V.C. Identification of medicinal mushrooms using computer vision and convolutional neural network. In Proceedings of the 2022 6th International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM), Medan, Indonesia, 22–23 November 2022; pp. 167–171. [CrossRef]
9. Martinez, J.; Escio, J.; Pellegrino, R. Identification of fish freshness through gill color using CNN and LAB image segmentation. In Proceedings of the 2023 IEEE 5th Eurasia Conference on IOT, Communication and Engineering (ECICE), Yunlin, Taiwan, 27–29 October 2023; pp. 563–567. [CrossRef]
10. Buenconsejo, L.T.; Linsangan, N.B. Detection and identification of abaca diseases using a convolutional neural network (CNN). In Proceedings of the TENCON 2021–2021 IEEE Region 10 Conference, Auckland, New Zealand, 7–10 December 2021; pp. 94–98. [CrossRef]
11. Morales, K.; Castillo, C.; Pellegrino, R. Identification of urban sounds with haptic feedback using Raspberry Pi and LSTM-SVM. In Proceedings of the 2024 IEEE International Conference on Innovative Computing and Artificial Intelligence Engineering Technologies (IICAIET), Kota Kinabalu, Malaysia, 26–28 August 2024; pp. 210–215. [CrossRef]
12. Diloy, R.; Juana, M.; Yumang, A. Coffee bean detection using Mask R-CNN. In Proceedings of the 2024 14th International Conference on Computer and Automation Engineering (ICCAE), Sydney, Australia, 12–14 March 2024; pp. 1–5. [CrossRef]
13. Villamor, I.V.; Apelo, M.L.D.; Ascan, D.A.C. Philippine coin detection system using Mask R-CNN algorithm. In Proceedings of the 2023 IEEE 15th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), Palawan, Coron, Philippines, 19–23 November 2023; pp. 1–6. [CrossRef]
14. Elizalde, D.Q.R.; Garcia, R.J.P.; Mitra, M.M.S.; Maramba, R.G. Wireless automated fire detection system on utility posts using ATmega328P. In Proceedings of the 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), Baguio City, Philippines, 29 November–2 December 2018; pp. 1–5. [CrossRef]
15. Calimag, A.D.R.; Padilla, D.A.; Manlises, C.O. Checkout system with object detection using NVIDIA Jetson Nano and Raspberry Pi. In Proceedings of the 2023 IEEE 5th Eurasia Conference on IOT, Communication and Engineering (ECICE), Yunlin, Taiwan, 27–29 October 2023; pp. 168–171. [CrossRef]
16. Assael, Y.M.; Shillingford, B.; Whiteson, S.; de Freitas, N. LipNet: End-to-end sentence-level lipreading. *arXiv*. 2016. Available online: https://arxiv.org/abs/1611.01599 (accessed on 4 October 2025).
17. Xue, F.; Yang, T.; Liu, K.; Hong, Z.; Cao, M.; Guo, D.; Hong, R. LCSNet: End-to-end lipreading with channel-aware feature selection. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 1–21. [CrossRef]
18. Al-Humaidan, N.; Prince, M. A classification of Arab ethnicity based on face image using deep learning approach. *IEEE Access* **2021**, *9*, 50755–50766. [CrossRef]