Indigenous People's Language Identification Using Machine Learning for Linguistic Preservation

Adrales, Lorelyn F.; Garingo, Joshua Razzi B.; Geraldez, Jan Anthony Q.; Taboada, Vene Lucille T.; Taladtad, Jelan Roy L.*

College of Engineering, Architecture and Technology, Notre Dame of Dadiangas University, General Santos City, Philippines

***Corresponding Author**

DOI: https://dx.doi.org/10.51584/IJRIAS.2025.1010000080

Received: 24 October 2025; Accepted: 30 October 2025; Published: 08 November 2025

ABSTRACT

Language is a fundamental aspect of human identity, deeply connected to geographical origins, cultural heritage, and social belonging. However, many indigenous languages across the world are gradually declining due to modernization, migration, and the growing influence of technology and global languages. The loss of these languages often leads to the disappearance of cultural values, oral traditions, and historical knowledge. This study explores the integration of machine learning techniques such as Long Short-Term Memory (LSTM), Yoon Kim's Convolutional Neural Network model, and TextConvoNet in developing a mobile text-to-text identification and translation application for Blaan dialects spoken in General Santos City, Polomolok, and Sarangani. The goal of the application is to aid in the preservation and revitalization of the Blaan language while providing an accessible platform for both native speakers and learners to understand, translate, and communicate in their local dialects.

To evaluate the usability and effectiveness of the application, User Acceptance Testing (UAT) was conducted among selected users. Data were collected through structured interviews, document analysis, and standardized evaluation tools to ensure comprehensive assessment and validation. Experimental results showed that the TextConvoNet model achieved the highest accuracy rate of 74.00 percent, surpassing the performance of both LSTM and CNN-based models. This demonstrates the model's efficiency in identifying and classifying Blaan dialects, highlighting its potential in the field of Natural Language Processing (NLP).

Future research should focus on expanding the dataset by collecting transcriptions from diverse age groups, locations, and communication contexts to improve model generalization and accuracy. Further refinement of the model's architecture and parameter tuning is also recommended to enhance dialect classification and translation capabilities. Moreover, integrating speech-to-text and text-to-speech functionalities could facilitate real-time translation, pronunciation learning, and accessibility for non-literate speakers, ensuring the continued preservation and appreciation of indigenous languages.

Keywords: Natural Language Processing (NLP), TextConvoNet, Yoon Kim, LSTM, Machine Learning

INTRODUCTION

Language is the foundation of human communication, enabling us to share thoughts, feelings, and more. Each country around the world has its own unique native languages. However, due to technological evolution, many languages are now under threat of extinction, which could result in the loss of cultural identity and traditional knowledge. Approximately 6,000 indigenous languages are recognized, and half of these, specifically 1,500, are at risk. One way to safeguard cultural heritages and indigenous languages is to integrate this knowledge into international science-policy-society frameworks, where diversity is recognized and contributes to strengthening Indigenous Peoples [1].

Currently, at least 182 languages are spoken in the Philippines [2], yet six languages are recognized as extinct [2][3][4]: Agta Villaviciosa, Agta Dicamay, Katabaga, Ayta Tayabas, Inagta Isarog, and Agta Sorsogon.

Notably, all these languages were once used by various Negrito ethnolinguistic communities. Lobel (2015) recounts that during his 2006 visit to one such community, he found no evidence of Black Filipino features among the inhabitants, who reported speaking only the Bikol Legaspi dialect and noted that their ancestors had spoken a different language [5].

Indigenous Leaders in Mindanao expressed that the non-use of Indigenous Language among the community will greatly influence the knowledge and wisdom affecting the younger generations. It is observed that the increase of misuse of global cell phones contributed to the decline of language use [6].

The study preserves indigenous languages and keeps pace with technological advancement by applying Machine Learning Techniques. It identifies and translates texts from different Blaan dialects using a mobile application developed by the researchers. Furthermore, the study provides a technological foundation for creating educational resources, translation tools, and digital archives which fosters a greater appreciation and understanding of regional linguistic diversity.

LITERATURE REVIEW

Language Identification

In the field of Natural Language Processing (NLP), language identification plays a crucial role in automatically determining the language used in a given text or document. This process is essential for tasks involving text analysis, as identifying the language is a prerequisite for further linguistic processing. A study examined the prediction of language using various machine learning algorithms, including Naïve Bayes, Logistic Regression, Decision Tree, and Random Forest [7]. The authors employed vectorization techniques to convert the text into a matrix format for analysis. This study compared the performance of these classification algorithms, highlighting the effectiveness of machine learning models in Blaan dialect identification based on various performance metrics.

Machine Learning

Machine Learning is a diversified subject of computer science and control science. It is widely used in various fields in which it learns patterns and relationships of complex data. Machine Learning has also the ability to adapt and generalize data models enabling it to compute complex processes. Moreover, there are three categories of Machine Learning, supervised learning, semi-supervised learning, and unsupervised learning [8].

Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs) are deep learning models effective for recognizing patterns in both visual and sequential data. Originally designed for image processing, they are now widely used in Natural Language Processing (NLP) tasks like text classification, sentiment analysis, and Blaan dialect identification. CNNs capture essential text features, enhancing accuracy in language tasks [9].

A recent advancement, TextConvoNet improves text classification by capturing both intra- and inter-sentence n-gram features using a two-dimensional convolution approach. In tests across five datasets, TextConvoNet outperformed traditional models in various performance metrics. This capability makes it a valuable tool for Blaan dialect identification and preserving Indigenous languages [9].

A significant contribution to the field of text classification is the model developed by Yoon Kim [10], in which he demonstrated the effectiveness of CNNs applied to sentence classification tasks. His work showed that simple CNN architectures could achieve competitive performance on multiple NLP benchmarks, including sentiment analysis and question classification. By utilizing pre-trained word embeddings, his model could capture relevant text representations and improve classification accuracy. Kim's methodology primarily relies on convolutional layers for feature extraction, utilizing various filter sizes to learn patterns effectively direct from the input text data. This approach laid the groundwork for later developments in CNN models, influencing subsequent architectures like TextConvoNet.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue X October 2025

Recurrent Neural Network (RNN)

Recurrent Neural Networks consist of interconnected nodes which are trained through the process of forward and backward propagation utilizing "binary cross-entropy loss" and optimized using the "Adam optimizer" to minimize the difference between predicted and actual labels. Under the variation of Recurrent Neural Network are:

LSTMs have proven highly effective in different activities, covering areas like understanding human language, forecasting sequences over time, and recognizing spoken language. It is mainly designed for processing sequential data of varying lengths [11].

METHODOLOGY

This chapter will present the research methodology employed in this study. It comprises the research design, selection of respondents, the instruments that will be utilized for research, the data collection process, and ethical considerations. Additionally, this chapter is composed of the outline of the data analysis methods, detailing the statistical formulas applied to validate the effectiveness of the algorithm.

Datasets

To identify a Blaan dialect, a dataset of text samples from the General Santos City, Polomolok, and Sarangani Blaan dialects was collected. The data was obtained from transcribed literature and interviews, ensuring a diverse representation of dialectal variations.

Table I. Blaan Texts by Category

Category	**Number of Texts**
Blaan Gensan	3038
Blaan Polomolok	2677
Blaan Sarangani	3320
Total	**9035**

Table I summarizes the number of texts gathered from each Blaan dialect. The largest portion of texts comes from Sarangani, with a total of 3320 texts. General Santos City follows with 3038 texts,while Polomolok has the smallest count at 2677 texts. In total, 9035 texts were gathered from all dialects.

Text Convo Net

TextConvoNet is a CNN-based architecture designed to capture both intra-sentence and inter-sentence n-gram dependencies in text data [9]. While conventional CNN models rely on 1D convolutional filters to extract intra-sentence features, TextConvoNet extends this by leveraging 2D multi-scale convolutional operations, enabling feature extraction across sentence boundaries. The model represents textual input as a paragraph-level embedding matrix, allowing for a richer representation of contextual relationships. Its architecture consists of multiple convolutional layers with varying kernel sizes, followed by ReLU activations, feature concatenation, and a classification layer.

To adapt TextConvoNet for multi-class classification, the researchers replaced the binary cross-entropy loss with sparse categorical cross-entropy and modified the activation function from sigmoid to softmax. This adjustment allows the model to output probability distributions over three classes rather than a binary decision. The

restructuring ensures effective distinction among the three dialects while leveraging a more efficient loss function for integer-labeled categorical data. Additionally, variations such as TextConvoNet_4 and TextConvoNet_6, differing in convolutional depth, were explored to assess their impact on classification performance [9].

Oversampling

In machine learning, class imbalance occurs when certain categories in a dataset have significantly more samples than others, leading to biased models that favor the majority class [12]. This imbalance can hinder a model's ability to accurately classify underrepresented categories, as it may learn to prioritize the dominant class while overlooking the minority ones. A common approach to address this issue is oversampling, which involves increasing the number of instances in the minority class by duplicating existing samples or generating synthetic ones. By balancing the dataset, oversampling improves model performance and ensures that all classes are adequately represented during training [13].

The researchers identified an imbalance in the dataset used for Blaan dialect identification, which could negatively impact the model's ability to classify underrepresented dialects accurately. To address this issue and develop a more effective classification model, they applied oversampling techniques to balance the dataset [12].

Regularization

Regularization is a technique in machine learning used to prevent overfitting, a common issue where models perform well on training data but fail to generalize to new, unseen data [14]. Overfitting often occurs when models become too complex, capturing noise and random fluctuations rather than meaningful patterns [15].

The researchers applied kernel regularization to enhance the model's ability to generalize across different Blaan dialects. This approach adds a penalty term to the loss function based on the magnitude of the kernel weights, discouraging excessively large weights and reducing the risk of overfitting. By incorporating kernel regularization, they aimed to develop a more robust and reliable classification model [16].

Class Weights

Class weights offer a complementary approach to RandomOverSampling (ROS) by adjusting the loss function to assign higher weights to the minority class, making the model more sensitive to its patterns [17]. Unlike oversampling, which modifies the dataset, class weighting directly influences learning by penalizing misclassification of underrepresented classes more heavily. In scikit-learn, classifiers like Logistic Regression and Random Forest use the class_weight parameter to adjust weights inversely proportional to class frequencies, improving precision and recall [18].

The researchers applied class weighting by computing dynamic weights, ensuring fair representation of all classes. The computed weights were converted into a dictionary and adjusted by boosting the most underrepresented class. This approach, combined with RandomOverSampling, helped mitigate class imbalance and improve classification performance.

Ethical Considerations

The study strictly adhered to ethical guidelines to protect the well-being of both the researchers and participants. Prior to data collection, ethical approval was obtained from the appropriate authorities, including formal authorization from the two local government units (LGUs) where the research was conducted. Informed consent was secured from all participants, ensuring their voluntary participation. Confidential information was securely handled, and all respondents' identities were kept anonymous to maintain privacy and confidentiality. The researchers also took careful measures to prevent any form of harm or distress throughout the research process, thereby ensuring the validity, reliability, and integrity of the results.

The researchers declare that there are no potential conflicts of interest, whether financial, personal, or professional, that could have influenced the conduct, analysis, or reporting of this study. All procedures and

findings were carried out objectively and independently, solely for academic and research purposes.

RESULTS AND DISCUSSION

Table II. Verb-Subject-Object (V-S-O) Pattern in Blaan Dialects

Dialect	**Verb**	**Subject**	**Object**
General Santos City	Salu	agu di	General Santos City.
Polomolok	Ngatu	agu di	General Santos City.
Sarangani	Gatu	agu di	General Santos City.
English (Direct translation)	Going	me to	General Santos City.
	(I am going to General Santos City)		

Table II presents the sentence structure analysis of the Blaan dialects spoken in General Santos City, Polomolok, and Sarangani. All three dialects consistently follow the Verb–Subject–Object (V-S-O) pattern, as illustrated in the example sentence "I am going to General Santos City." This structural consistency aligns with previous linguistic findings, which confirmed that Blaan sentences predominantly use the V-S-O construction across the Soccsksargen Region [19]. While this pattern is comparable to Tagalog, it contrasts with the Subject–Verb–Object structure typical of English [20]. The results indicate that the V-S-O order is a fundamental syntactic feature of the Blaan language. Recognizing this shared sentence structure is essential for developing accurate translation models, instructional materials, and linguistic documentation that support the preservation and continued understanding of the Blaan language.

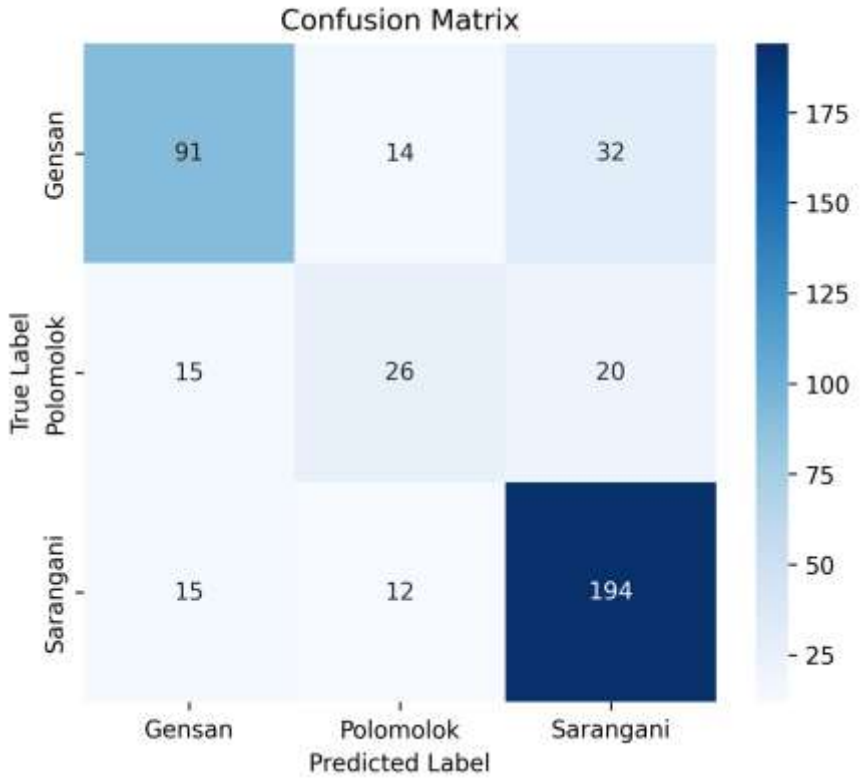
Table III. Blaan Dialects Classification Report

	Precision	**Recall**	**F1-Score**	**Support**
Blaan Gensan	0.75	0.66	0.71	137
Blaan Polomolok	0.50	0.43	0.46	61
Blaan Sarangani	0.79	0.88	0.83	221
Accuracy			0.74	419
Macro Average	0.68	0.66	0.67	419
Weighted Average	0.73	0.74	0.74	419

Table III presents the classification report for TextConvoNet's performance in identifying Blaan dialects. The model achieved an overall accuracy of 74%, meaning it correctly classified most samples. Among the dialects, Blaan Sarangani had the highest recall (0.88), indicating that the majority of Sarangani samples were correctly identified. It also had the highest F1-score (0.83), showing a strong balance between precision and recall. Blaan Gensan had a recall of 0.66, meaning 66% of its samples were correctly classified, with an F1-score of 0.71. Blaan Polomolok, however, had the lowest recall (0.43) and F1-score (0.46), indicating that the model struggled the most with distinguishing Polomolok from the other dialects. The macro average F1-score of 0.67 highlights

the imbalance in performance across dialects, while the weighted average of 0.74 suggests that, overall, the model performs well but has room for improvement, particularly in differentiating Polomolok from Gensan and Sarangani.

Figure 1. Text Convo Net Confusion Matrix



The figure above presents the confusion matrix for TextConvoNet's classification of Blaan texts. It correctly identified 91 Gensan, 26 Polomolok, and 194 Sarangani samples. However, misclassifications occurred, with some Gensan samples predicted as Polomolok or Sarangani, and similar errors affecting Polomolok and Sarangani. The model performed well overall but struggled with Polomolok, indicating a need for further improvements.

Table IV. Comparison of Model Performance

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
TextConvoNet	74.00%	68.00%	65.66%	66.66%
Yoon Kim	70.00%	65.66%	69.00%	65.66%
LSTM	70.00%	67.66%	70.33%	67.00%

Table IV presents a comparison of model performance based on accuracy, precision, recall, and F1-score. Among the three models, TextConvoNet achieved the highest accuracy at 74.00%, with a precision of 68.00%, recall of 65.66%, and an F1-score of 66.66%. Yoon Kim's CNN-based model, on the other hand, attained an accuracy of 70.00%, with a precision of 65.66%, recall of 69.00%, and an F1-score of 65.66%. Similarly, the LSTM model also recorded an accuracy of 70.00% but demonstrated a slightly higher precision of 67.66%, recall of 70.33%, and an F1-score of 67.00%. Since the researchers' mobile application identifies the dialect of a given input text, TextConvoNet is considered the most suitable model due to its higher overall accuracy. While LSTM offers a balanced trade-off between precision and recall, TextConvoNet provides more reliable predictions for individual text inputs, making it the preferred choice for the application.

The results indicate that TextConvoNet effectively classifies Blaan Sarangani texts but struggles with Blaan Polomolok, as seen in its low recall (0.43) and F1-score (0.46). This discrepancy could be due to data distribution and linguistic similarities between dialects. Sarangani had the highest support (221 samples), which may have contributed to the model's ability to learn its features more effectively. In contrast, Polomolok had the lowest support (61 samples), potentially leading to underrepresentation in the training process. The model's confusion matrix further supports this, showing that a significant portion of Polomolok samples were misclassified as either Gensan or Sarangani. These errors suggest that Polomolok shares overlapping linguistic characteristics with the other dialects, making it harder for the model to differentiate.

Additionally, the macro average F1-score of 0.67 reveals an imbalance in classification performance across dialects, reinforcing the idea that the model may require further refinement in feature extraction or additional training data for Polomolok to improve its accuracy. The weighted average of 0.74, however, suggests that while there is room for improvement, the model remains fairly robust overall.

The model comparison indicates that TextConvoNet outperforms both Yoon Kim's model and LSTM in terms of overall accuracy (74.00%), making it the most reliable model for dialect identification in the researchers' mobile application. Although LSTM demonstrates a slightly higher recall (70.33%) compared to TextConvoNet (65.66%), the latter provides a better balance of precision and recall, as reflected in its F1-score of 66.66%. Yoon Kim's model, while showing comparable recall (69.00%) to LSTM, has a lower precision and an identical F1-score (65.66%), indicating a less consistent performance. The higher accuracy of TextConvoNet suggests that it is more effective in correctly identifying dialects on a per-text basis, aligning well with the intended functionality of the application.

The findings have important implications for the development of automated Blaan dialect identification and translation systems. The model's strong performance on Blaan Sarangani suggests that it could be effectively deployed for dialect identification, particularly in areas where Blaan Sarangani is predominantly spoken. However, the misclassification of Blaan Polomolok highlights the need for data augmentation, improved preprocessing techniques, or more advanced architectures to enhance differentiation between dialects.

Text-to-Text Mobile Application for Blaan Dialect Identification and Translation

Figure 2. Privacy Notice of the Mobile Application

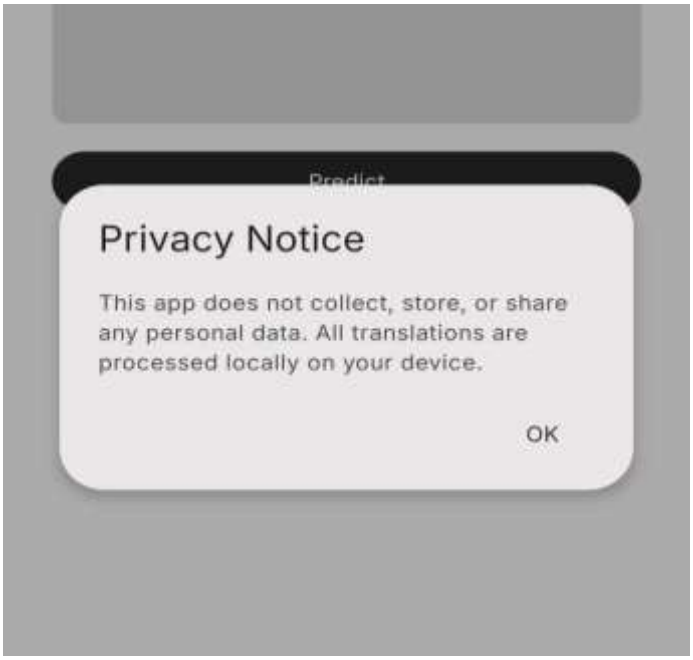


Figure 2 shows the Privacy Notice displayed when the application is launched for the first time. This notice informs users about the app's commitment to privacy and ensures transparency by clarifying that no personal data is collected, stored, or shared.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue X October 2025

Figure 3. Dialect Identification Feature of the Mobile Application

Figure 3 illustrates the dialect identification feature, where users input a Blaan text, and the application predicts the most likely dialect along with confidence scores. The results are displayed in both text and graphical formats, providing a clear breakdown of the classification probabilities for Gensan, Polomolok, and Sarangani dialects.

Figure 4. Translation Feature of the Mobile Application

Figure 4 showcases the translation feature, where users can input either a Blaan text or an English sentence. The application allows users to select the appropriate Blaan dialect before translating. Upon clicking the Translate button, the system generates the corresponding translation from Blaan to English or from English to Blaan, displaying the output below the button. This feature enables users to understand Blaan phrases and also construct Blaan translations from English based on the selected dialect.

The results show that the mobile application can accurately identify Blaan dialects and translate texts while preserving their meaning. The dialect identification feature effectively distinguishes between the three Blaan dialects based on linguistic patterns from the dataset.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue X October 2025

The translation feature provides reliable translations by using predefined translations from collected data, ensuring that key contextual meanings are maintained. Compared to general translation tools that do not support Blaan, this application offers a more specialized and accurate solution.

Some challenges, such as limited data for certain dialects and variations in spelling, affected accuracy. However, these issues were minimized through careful data processing and verification.

The study highlights the importance of using technology to support the preservation and accessibility of the Blaan language. The app serves as a valuable tool for Blaan speakers, students, and researchers by providing an easy way to classify dialects and access translations.

In education, the app can help teach Blaan dialects, while in daily life, it can assist in communication between Blaan speakers and non-speakers, particularly in schools, government offices, and healthcare services.

Future improvements could include expanding the translation database, refining dialect classification accuracy, and integrating voice recognition. These updates would further enhance the app's role in language preservation and cross-cultural communication.

User Acceptance Testing

Table V. Detailed UAT Result of the Program Design

Indicator	Weighted Mean	Verbal Interpretation
1. The user interface is visually appealing.	4.30	Excellent
2. The application is easy to navigate.	4.83	Excellent
3. The design elements throughout the system are consistent.	4.80	Excellent
4. The application's layout is intuitive.	4.70	Excellent
5. The application loads quickly and has little to no delay when in use.	4.57	Excellent

Table V presents the detailed UAT results of the Program Design, highlighting the mobile application's excellent design and usability. Users find the interface visually appealing with a score of 4.30, easy to navigate with 4.83, and consistently designed with 4.80. The layout is intuitive, scoring 4.70, while the app performs smoothly with minimal delays, receiving 4.57. These results confirm that the application is highly functional, user-friendly, and well-optimized.

Table VI. Detailed UAT Result of the Acceptability

Indicator	Weighted Mean	Verbal Interpretation
1. The application can be favorably received and used.	4.63	Very Acceptable
2. I am comfortable with the level of guidance provided for the Blaan dialect identification process.	4.73	Very Acceptable
3. The mobile application's privacy measures, including data protection and anonymity, are clearly communicated and satisfactory.	4.83	Very Acceptable

4. The overall user experience of the mobile application meets my expectations and is enjoyable.	4.70	Very Acceptable
5. I would recommend this application to others who are interested in Blaan dialect identification and translation	4.87	Very Acceptable

Table VI presents the detailed UAT results of the Acceptability, showing that the mobile application is well-received, with all criteria rated "Very Acceptable." Users find the application favorable for use, scoring 4.63, and are comfortable with the level of guidance provided for the Blaan dialect identification process, which received 4.73. Privacy measures, including data protection and anonymity, are clearly communicated and satisfactory, scoring 4.83. The overall user experience is enjoyable, with a score of 4.70, and the application is highly recommended for those interested in Blaan dialect identification and translation, receiving the highest rating of 4.87. These results confirm the application's wide acceptance, user-friendliness, and effectiveness.

Table VII. User Acceptance Test Summary

Indicator	**Overall Weighted Mean**	**Verbal Interpretation**
1. Program Design	4.64	Excellent
2. Acceptability	4.75	Very Acceptable

Table VII presents the User Acceptance Test Summary, confirming that the mobile application performs excellently in both program design and acceptability. The program design received an overall weighted mean of 4.64, rated as "Excellent," indicating that users find the interface, navigation, and overall usability highly satisfactory. Acceptability scored 4.75, rated as "Very Acceptable," showing that users positively received the application, found it user-friendly, and would recommend it to others. Overall, these results affirm that the application meets user expectations and functions effectively in both design and usability.

Users provided positive feedback on the application, highlighting its usefulness, accessibility, and ease of use, particularly for those learning the Blaan dialect. They appreciated its role in improving communication and understanding, with some suggesting the addition of "bantas" or punctuation marks. Others recognized the app's potential and recommended minor design enhancements, such as automatically hiding the keyboard after pressing the translate button and adjusting font styles and colors for better readability. Overall, the feedback reinforces the application's effectiveness and usability, with minor refinements suggested to further enhance the user experience.

The User Acceptance Test results validate the mobile application's effectiveness in both design and functionality. The program design received an "Excellent" rating with a weighted mean of 4.64, confirming that users find the interface visually appealing, easy to navigate, and well-structured. Additionally, the acceptability of the application was rated "Very Acceptable" with a mean score of 4.75, highlighting its accessibility, privacy measures, and positive reception among users. The overwhelmingly favorable ratings suggest that the application meets user expectations and performs reliably in facilitating Blaan dialect identification and translation. Furthermore, qualitative feedback reinforces these findings, with users acknowledging the app's usefulness, especially for Indigenous People and students, while only minor refinements were suggested for an even better experience.

The findings indicate that the mobile application is a highly effective tool for Blaan dialect identification and translation, demonstrating strong usability and acceptability. The high user satisfaction suggests that the app can serve as a valuable resource for Indigenous communities, educators, and language learners. Additionally, the feedback on minor UI improvements such as better font styles, keyboard behavior, and punctuation additions provides actionable insights for enhancing user experience. Addressing these refinements will further optimize accessibility, engagement, and usability, ensuring that the application continues to serve its purpose effectively

while promoting the preservation and understanding of the Blaan language.

CONCLUSIONS

Blaan dialects represent an essential component of cultural identity among communities in General Santos City, Polomolok, and Sarangani, underscoring the importance of developing technological interventions for language preservation. This study employed the TextConvoNet model to classify and translate dialectal variations of the Blaan language and confirmed the consistent use of a Verb–Subject–Object (V-S-O) sentence structure, supporting existing linguistic analyses and informing the development of structured translation approaches. The application of oversampling, class weighting, and kernel regularization contributed to improved model training by addressing class imbalance and mitigating overfitting. The model achieved a validation accuracy of 97.74%, indicating strong classification capabilities; however, the test accuracy of 74.00% suggests the need for further refinement to improve generalization, particularly in distinguishing the Polomolok dialect from other variants. The successful integration of the trained model into a functional mobile application demonstrates its practicality for real-world use, with User Acceptance Testing confirming high functionality, usability, and user satisfaction. Overall, the study highlights the potential of deep learning in supporting indigenous language preservation and provides a foundation for future research aimed at enhancing model robustness, translation accuracy, and accessibility to promote sustained linguistic and cultural continuity.

RECOMMENDATIONS

Based on the findings of this study, several directions for further improvement and broader application are recommended. The system may be expanded to include other indigenous languages and dialects to promote wider linguistic preservation and strengthen cultural inclusivity. Collecting additional data from varied sources, such as oral conversations, recorded community interactions, written literature, and digital archives, would provide a more diverse linguistic base and capture differences across age groups, locations, and social contexts. This expanded dataset may also support improved recognition of dialectal features and language variations. Further refinement of the model can be achieved through experimentation with alternative neural architectures, parameter tuning, and more extensive evaluation to enhance classification performance. Integrating speech-to-text and text-to-speech capabilities would support real-time pronunciation learning and broaden accessibility for both native speakers and new learners, particularly those who are more familiar with spoken language than written forms. Additionally, developing a translation component that can generate context-aware translations rather than relying primarily on direct text matches would increase flexibility and practical usefulness in real communication scenarios. These improvements can contribute to more effective language preservation and support the continued intergenerational transmission of indigenous linguistic knowledge.

REFERENCES

1. UNESCO Ad Hoc Expert Group on Endangered Languages. (2003, March 10–12). Language vitality and endangerment. International Expert Meeting on the UNESCO Programme Safeguarding of Endangered Languages, Paris, France. https://ich.unesco.org/doc/src/00120-EN.pdf
2. Lewis, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2015). Ethnologue: Languages of the world (18th ed.). SIL International. http://www.ethnologue.com
3. Komisyon sa Wikang Filipino. (2018). Kapasiyahan ng kalupunan ng mga komisyoner blg. 18-33 serye 2018.
4. Headland, T. N. (2003). Thirty endangered languages in the Philippines. Work Papers of the Summer Institute of Linguistics, University of North Dakota Session, 47(1). https://commons.und.edu/sil-work-papers/vol47/iss1/1
5. Pelila, J. R. O., Ayao-ao, S. L., & Casiano, M. B. (2023). If these languages could talk: The extinct languages of the Philippines. International Journal of Multidisciplinary Research and Publications (IJMRAP), 6(3), 127–134.
6. Villaluz, RSCJ, PhD., Geraldine D.; Tagalog, EdD, ISRM, Rita May P.; and Saway, Aduna L. Bai (2023). Engaging Indigenous Community Towards a Talaandig Language Learning and Cultural

Sustainability. ASEAN Journal of Community Engagement, 7(2), 129-150. https://doi.org/10.7454/ajce.v7i2.1227

7. Abinaya, N., Jayadharshini, P., Priyanka, S., Keerthika, S., & Santhiya, S. (2023). Identification of language from multi-lingual dataset using classification algorithms. Journal of Physics: Conference Series, 2664(1), 012009. https://doi.org/10.1088/1742-6596/2664/1/012009
8. Zhou, J., & Huang, T. (2023). Application of machine learning algorithm in electronic book database management system. SN Applied Sciences, 5(11), 287. https://doi.org/10.1007/s42452-023-05508-3
9. Soni, S., Chouhan, S. S., & Rathore, S. S. (2023). TextConvoNet: a convolutional neural network based architecture for text classification. Applied Intelligence, 53(11), 14249-14268. https://doi.org/10.1007/s10489-022-04221-9
10. Kim**,** Y. (2014). Convolutional Neural Networks for Sentence Classification. arXiv preprint arXiv:1408.5882. https://arxiv.org/abs/1408.5882
11. Bahad, P., Saxena, P., & Kamal, R. (2019). Fake News Detection using Bi-directional LSTM-Recurrent Neural Network. Procedia Computer Science, 165, 74-82. [DOI: 10.1016/j.procs.2020.01.072]
12. Shivani Rana, Rakesh Kanji, & Shruti Jain. (2024). Comprehensive Analysis of Oversampling Techniques for Addressing Class Imbalance Employing Machine Learning Models. In Recent Advances in Computer Science and Communications. https://www.semanticscholar.org/paper/cbc2b83a63befe1d8631828d3fa7365c087579f5
13. Elsobky, A. Keshk, & M. Malhat. (2023). A Comparative Study for Different Resampling Techniques for Imbalanced datasets. In IJCI. International Journal of Computers and Information. https://www.semanticscholar.org/paper/d046d080e8caf355e5af48ae6dd6bdb14ec4cec3
14. Gouranga Jha. (2024). Popular Machine Learning Models Prone to Overfitting and Why It ... https://medium.com/@post.gourang/popular-machine-learning-models-prone-to-overfitting-and-why-it-happens-8050e9c3a944
15. Chenlei Fang. (2020). 4 – The Overfitting Iceberg – Machine Learning Blog | ML@CMU. https://blog.ml.cmu.edu/2020/08/31/4-overfitting/
16. Darshan M. (2022). How do Kernel Regularizers work with neural networks? https://analyticsindiamag.com/ai-trends/kernel-regularizers-with-neural-networks/
17. Olamendy, J. (2024). Practical ML: Addressing Class Imbalance | by Juan C Olamendy. https://medium.com/@juanc.olamendy/practical-ml-addressing-class-imbalance-25c4f1b97ee3
18. Himanshi Singh. (2024). 10 Techniques to Solve Imbalanced Classes in Machine Learning. https://www.analyticsvidhya.com/articles/class-imbalance-in-machine-learning/
19. Lobel, J. W. (2015). Philippine and North Bornean languages: Issues in description, subgrouping, and reconstruction (Doctoral dissertation, University of Hawai'i Manoa). http://www.ling.hawaii.edu/graduate/Dissertations/JasonLobelFinal.pdf
20. Glossika Content Team. (2023). Tagalog Language Overview: A Bigger Picture For Beginners. https://ai.glossika.com/blog/tagalog-language-overview