Exploratory data analysis and forecasting of dengue outbreaks in Pangasinan using the ARIMA model

Patrick Mole, Thelma Palaoag
College of Information Technology and Computer Science, University of the Cordilleras, Baguio, Philippines

Article Info

Article history:

Received Dec 2, 2024
Revised Jun 30, 2025
Accepted Aug 6, 2025

Keywords:

ARIMA model
Dengue forecasting
Exploratory data analysis
Predictive model
Time series analysis

ABSTRACT

Dengue fever remains a critical public health concern in tropical countries like the Philippines, with Pangasinan frequently experiencing outbreaks due to favorable environmental conditions for mosquito breeding. Despite ongoing efforts to control the disease, the absence of a reliable forecasting tool limits the ability of health authorities to implement proactive measures. This study developed a forecasting model using the autoregressive integrated moving average (ARIMA) technique, following an initial exploratory data analysis (EDA) to identify trends and patterns in historical dengue case data from 2019 to 2024. The ARIMA model was trained and validated using historical data, capturing seasonal variations and projecting future dengue outbreaks. The evaluation metrics, including mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE), indicated that the model achieved an accuracy of approximately 78.3%, suggesting reasonable predictive capability. Forecasts for the year 2025 indicate a potential rise in dengue cases, particularly during peak seasons, aligning with observed historical trends. These predictions offer valuable insights for local health authorities, enabling them to plan targeted interventions, allocate resources efficiently, and mitigate the impact of future outbreaks. The study demonstrates the practical application of time series analysis in public health forecasting and provides a proactive tool tailored for the needs of Pangasinan.

Corresponding Author:

Patrick Mole
College of Information Technology and Computer Science, University of the Cordilleras
Gov. Pack Rd., Baguio, Philippines
Email: pvm8598@students.uc-bcf.edu.ph

1. INTRODUCTION

Dengue fever remains a significant public health challenge in many tropical countries around the world, where environmental conditions provide an ideal setting for the proliferation and spread of Aedes mosquitoes, the primary vectors responsible for transmitting the disease. It is prevalent in 126 countries, placing 3.9 billion people, or 48% of the global population, at risk, with an estimated 390 million infections and up to 36,000 deaths occurring annually [1], [2]. Dengue infection can range from being asymptomatic to causing mild fever, but in some cases, it can escalate to a life-threatening condition known as dengue hemorrhagic fever [3].

In the Philippines, dengue remains a pressing health concern. The Department of Health (DOH) consistently reports high numbers of dengue cases, straining healthcare systems across the country [4]. As of October 2024, the DOH reports a cumulative total of 269,467 dengue cases, reflecting an 82% rise compared to the 147,678 cases recorded during the same period last year [5]. Meanwhile in Pangasinan, a densely populated province in the northern Philippines, dengue cases have surged dramatically, mirroring national trends. The Pangasinan Provincial Health Office (PHO) reports a 52% increase in dengue cases in 2024, leading to 14 deaths as of mid-year [6].

Despite ongoing public health efforts to control the spread of dengue, the unpredictable nature of outbreaks poses significant challenges to the timely implementation of effective interventions. One of the key challenges in managing dengue outbreaks is the absence of an effective system for predicting future outbreaks. Current response strategies tend to be reactive rather than proactive, often leading to delayed mobilization of resources. Without a reliable forecasting model, health officials are left to rely on intuition on historical data, which limits the ability to implement preventive measures in high-risk areas before an outbreak occurs. This reactive approach not only results in higher infection rates but also increases the burden on healthcare systems and communities.

Recognizing the critical need for early intervention and addressing existing challenges in dengue outbreak control, this study seeks to develop and validate a forecasting model based on historical dengue case data to accurately predict future outbreaks in Pangasinan, thereby providing a proactive tool for public health planning and intervention. Specifically, this study aims to address the lack of predictive systems by developing autoregressive integrated moving average (ARIMA) model, exploring to what extent it can be effectively applied to historical dengue case data of Pangasinan.

The ARIMA model, a widely used statistical method for time series analysis [7], is particularly suited to modeling dengue incidence trends based on historical case data [8]. In recent years, numerous studies have employed the ARIMA model to forecast other range of infectious diseases globally, including influenza [9], and COVID-19 [10], highlighting its effectiveness in capturing seasonal patterns and predicting future outbreaks with reasonable accuracy. However, while the ARIMA model has been applied successfully in predicting infectious diseases globally, there is a gap in its localized application for specific regions like Pangasinan. Many existing studies have focused on national or broader regional forecasts, overlooking the importance of localized prediction models that consider region-specific trends. Addressing this gap can help tailor public health responses to the specific needs of local communities.

The significance of this study lies in its potential to enhance the capacity of public health authorities to respond to dengue outbreaks proactively. By forecasting possible dengue surges, local officials can allocate resources more efficiently, target high-risk areas for preventive interventions, and ultimately reduce the spread of the disease. This study shall not only contribute to the growing body of research on dengue forecasting but also offer a practical solution tailored to the needs of Pangasinan, where dengue continues to threaten public health.

2. RESEARCH METHOD

This research study on forecasting dengue outbreaks using a predictive model followed a structured methodology. In the model selection phase, the researchers conducted an extensive review of recent related literature to identify the most suitable model for predicting dengue outbreaks. Relevant studies on time series forecasting methods were gathered and analyzed, focusing on models previously applied in the context of infectious disease prediction. Alternative models, such as seasonal autoregressive integrated moving average (SARIMA) and other machine learning approaches like long short-term memory (LSTM) networks, were considered. However, ARIMA was selected due to its interpretability, simplicity, and reliability, particularly for short-term dengue case predictions [11]. The decision to use ARIMA was guided by its ability to effectively model non-stationary time series data while capturing trends and seasonal variations. While SARIMA offers enhancements for explicitly handling seasonality, the seasonal component in dengue data for Pangasinan was adequately addressed using preprocessing and ARIMA parameters. Furthermore, machine learning models like LSTM, though powerful, require extensive computational resources, larger datasets, and are often less interpretable for non-technical stakeholders [12]. To effectively apply the selected ARIMA model, the following methodologies were employed to systematically develop, assess, and validate its predictive capabilities.

2.1. Data collection and preprocessing

The historical dengue case data used in this study were obtained from the Provincial Epidemiology and Surveillance Unit (PESU) of Pangasinan, covering the period from 2019 to 2024. The dataset includes monthly dengue case counts from various municipalities in Pangasinan. The dataset was subjected to thorough preprocessing to ensure data quality and consistency. Missing values were examined and handled using the mean-fill technique, as recommended by Kamalov [13], to maintain temporal continuity and minimize bias in the dataset. Duplicate records were identified and removed to prevent data repetition. To ensure uniformity in data representation, municipality names were standardized by correcting typographical errors and ensuring consistent spelling. Redundant columns, such as administrative notes or metadata not relevant to the analysis, were eliminated to streamline the dataset and improve computational efficiency.

Additionally, the dataset was reformatted into a structured time-series format, with dengue case counts indexed by month and year to facilitate accurate temporal analysis. This restructuring ensured that each observation followed a consistent chronological order, enabling precise trend detection and forecasting.

2.2. ARIMA model training and evaluation

The dataset was divided into 70% training data and 30% testing data to train and evaluate the model. This division adheres to established best practices in time-series forecasting, ensuring that a substantial portion of the data is used for model training while retaining a sufficient testing set for accurate evaluation and validation [14], [15]. Using at least two-thirds of the data for training ensures the model learns stable temporal patterns, while the remaining data allows for effective performance evaluation. The ARIMA model then was trained using the training dataset, enabling it to learn historical dengue case patterns. The model was implemented in R language using the forecast and tseries packages, which provide robust tools for time-series analysis. The forecast package was used for ARIMA model estimation and prediction, while the tseries package facilitated statistical testing and preprocessing.

To determine stationarity, the augmented dickey-fuller (ADF) test was performed, ensuring the data met ARIMA's assumptions. The model parameters were identified using autocorrelation function (ACF) and partial autocorrelation function (PACF) plots, while the akaike information criterion (AIC) was utilized to optimize the selection of parameters that minimized forecasting error.

2.2.1. ADF test for stationarity

To determine whether the time series data were stationary, the ADF test was performed. The ADF test follows the (1):

$$\Delta Y_t = \alpha + \beta t + \gamma Y_{t-1} + \sum_{i=1}^{p} \delta i \, \Delta Y_{t-i} + \epsilon t \quad (1)$$

Where Yt is the time series, ΔYt is the first-differenced value, t is time, α is a constant, β is the coefficient for trend, $\gamma Yt-1$ represents the lagged value, δi represents the lagged differenced terms, and ϵt is the white noise error term. The decision criteria for stationarity were based on the test statistic and p-value: If the test statistic is more negative than the critical value, the series is stationary. Thus, if the p-value is less than 0.05, the null hypothesis (presence of a unit root) is rejected, confirming stationarity.

2.2.2. Model selection using AIC

To identify the best-fitting ARIMA model, different model configurations were tested, and their AIC values were compared. AIC is calculated as (2):

$$AIC = 2k - 2ln(L) \quad (2)$$

Where: k is the number of estimated parameters, and L is the likelihood function. A lower AIC value indicates a better model fit.

Moreover, the model's performance was assessed using multiple statistical metrics, including mean absolute error (MAE) to measure the average magnitude of errors in forecasts, root mean squared error (RMSE) to capture large errors by penalizing higher deviations, and mean absolute percentage error (MAPE) to evaluate the model's accuracy by comparing predicted and actual values. Additionally, mean absolute scaled error (MASE) was used to assess relative accuracy compared to a naïve forecasting model, while the ACF was analyzed to check for patterns in prediction errors.

2.3. Utilization of the ARIMA model

In the final phase, the trained and validated ARIMA model was utilized to generate forecasts of dengue cases in Pangasinan. Diagnostic checks were conducted to ensure the reliability of the forecast, including residual analysis to confirm the absence of autocorrelation in errors, which ensures that the model's predictions are independent and accurate. These forecasting methods provide a structured approach to anticipating dengue outbreaks and serve as a foundation for proactive public health planning. By assessing the residuals and validating the model's assumptions, the model's performance was further confirmed, ensuring its suitability for making accurate predictions. The model was then applied to predict which municipalities are likely to experience dengue outbreaks and to identify the specific months when the province of Pangasinan may be at higher risk for the year 2025. The forecasts were visualized through time series plots, highlighting potential peak periods and enabling a clearer understanding of the anticipated trends in dengue incidence.

3. RESULTS AND DISCUSSION

The dataset used in this analysis consists of monthly dengue case records from various municipalities in Pangasinan, including four main variables: Month, Year, Municipality, and Dengue Cases. These variables provide a time-based breakdown of dengue cases, allowing for both temporal and spatial analysis of outbreak patterns. The results of this study are presented in two parts: (1) exploratory data analysis (EDA) of dengue data in Pangasinan, which examines historical trends and patterns in the dataset before proceeding to model development, and (2) Implementation of the ARIMA model, which focuses on the development, evaluation, and forecasting of dengue outbreaks using time-series modeling.

3.1. Exploratory data analysis of dengue data in Pangasinan

As part of this analysis, a box plot was generated to compare the yearly distribution of dengue cases across municipalities in Pangasinan. Figure 1 presents this visualization, illustrating variations in dengue incidence from 2019 to 2024. This plot highlights the spread of cases each year, identifies potential outliers, and provides insights into the effectiveness of dengue control measures over time. By examining these distributions, the analysis offers a clearer understanding of annual trends and the extent of dengue outbreaks within the province.

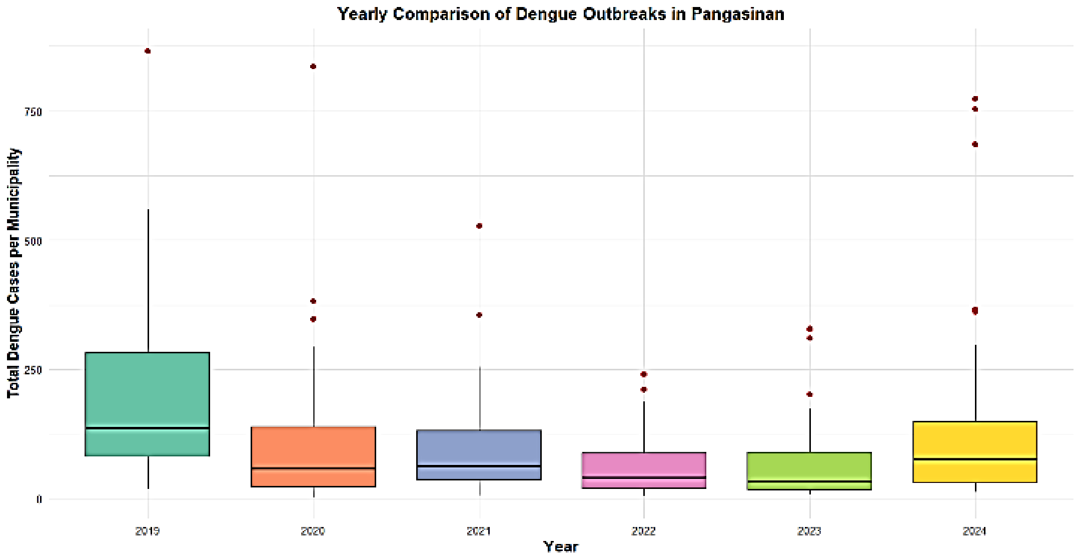

Figure 1. Comparison of annual dengue cases in Pangasinan from 2019 to 2024

The box plot in Figure 1 provides an overview of the yearly dengue case distribution in Pangasinan from 2019 to 2024, showcasing the variations in case counts across different years. The results indicate that 2019 experienced the highest median number of dengue cases, with a wider interquartile range and a greater spread, suggesting that dengue outbreaks were more severe and varied significantly during that year. In contrast, the median number of cases declined steadily from 2020 to 2023, reflecting a potential improvement in dengue control measures. The narrower interquartile ranges in 2021, 2022, and 2023 suggest more consistent and lower case number. However, in 2024, there is a noticeable increase in the median number of cases, along with a wider spread, indicating a possible resurgence of dengue outbreaks. The presence of outliers, represented by red dots, suggests sporadic surges in dengue cases during certain years, potentially indicating localized outbreaks. Notably, 2021 and 2022 have fewer extreme values, indicating more uniform case distribution. The sharp decline in cases from 2019 to 2023 suggests that control efforts may have been effective; however, the increase observed in 2024 underscores the need for continued vigilance and intervention. These findings emphasize the fluctuating nature of dengue outbreaks and the importance of sustained public health measures. However, while Figure 1 effectively captures overall trends at the provincial level, it does not provide insights into which municipalities contributed the most to these outbreaks or when peak dengue transmission occurred.

To address these gaps, Figure 2 presents a heatmap illustrating the spatial and temporal distribution of dengue cases across municipalities in Pangasinan. This figure allows for a more detailed analysis by identifying high-burden areas and seasonal patterns.

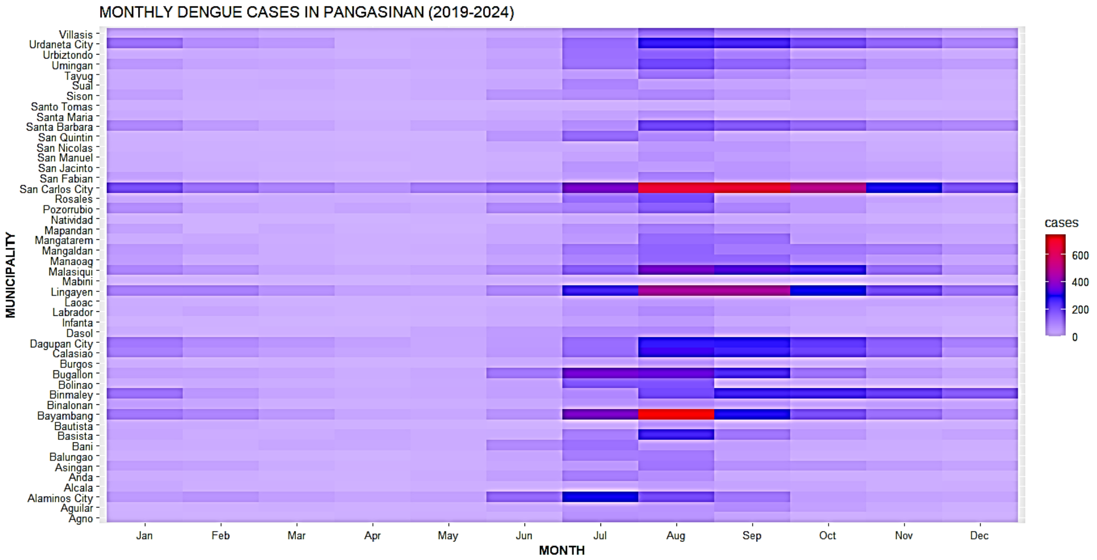

Figure 2. Heatmap of spatial distribution of dengue cases in Pangasinan from 2019 to 2024

The heatmap in Figure 2 provides a detailed visualization of the temporal and spatial distribution of dengue cases across various municipalities in Pangasinan from 2019 to 2024. The intensity of the color gradient represents the severity of dengue outbreaks, with brightest red color indicating higher case concentrations. A clear seasonal trend emerges, with dengue cases peaking predominantly between July and October each year. This pattern strongly aligns with the well-documented peak dengue season in the Philippines, which coincides with the country's rainy season from June to October [16]. During this period, the increased rainfall contributes to the accumulation of stagnant water in various environments, such as containers, drains, and waterlogged areas, providing optimal breeding grounds for Aedes aegypti, the primary mosquito vectors responsible for dengue transmission. Moreover, the heatmap reveals that dengue outbreaks are not uniformly distributed across municipalities, with certain areas experiencing persistently high case counts throughout the study period. These municipalities may have environmental conditions, population density, or sanitation challenges that contribute to sustained dengue transmission. The recurring peak in cases during the rainy season suggests that dengue outbreaks in Pangasinan follow a predictable annual cycle, emphasizing the critical need for proactive control measures before the onset of the rainy season.

While the EDA provides valuable insights into the historical trends and variations of dengue outbreaks in Pangasinan, it is clear that relying on these trends alone is insufficient for proactive public health management. The seasonal peaks, along with the spatial distribution of cases across municipalities, highlight the need for a forecasting tool that can anticipate future outbreaks and guide timely interventions. The observed fluctuations in the data, underscore the importance of developing a model that can predict these cyclical events with a degree of accuracy. Therefore, the ARIMA model was developed and applied to forecast future dengue outbreaks in Pangasinan. The following sections present the ARIMA model implementation, covering its training, evaluation, and use in forecasting future outbreaks.

3.2. Implementation of ARIMA model in dengue data of Pangasinan

Time series models such as ARIMA, SARIMA, and machine learning techniques have been widely used to predict infectious diseases; however, their application to localized dengue forecasting, especially for specific regions like Pangasinan, has been limited. Most previous studies have focused on national or broader regional predictions, often overlooking the distinct patterns that may exist at the municipal level. To address this gap, the ARIMA model was chosen and employs due to its ability to capture seasonal patterns and temporal dependencies, offering a more tailored and localized forecast for Pangasinan's unique epidemiological context.

3.2.1. ARIMA model training and testing

To build a predictive model, the dataset was divided into two segments: a training set and a testing set. Following the standard approach, the 70% of the data was allocated for training, and 30% for testing. The training set was used to develop and fit the ARIMA model, allowing it to learn historical trends and seasonal patterns. The remaining 30% served as the testing set, enabling the evaluation of the model's performance and forecast accuracy. Consequently, after training the ARIMA model, it was necessary to ensure that the time series data were stationary, as stationarity is a fundamental assumption for ARIMA modeling. To verify this, the ADF test was conducted on the dengue case data. The ADF test assesses whether a time series contains a unit root, indicating non-stationarity. The null hypothesis (H_0) states that the series is non-stationary, while the alternative hypothesis (H_1) suggests that the series is stationary. The results of the ADF test for the dengue case dataset are presented in Table 1.

Since the test statistic (-3.21) is more negative than the critical value (-2.89) and the p-value (0.021) is below the significance level (0.05), the null hypothesis was rejected. This confirms that the dengue case dataset is stationary, meaning it does not require additional differencing before model training. Once stationarity was confirmed, the next step involved selecting the best ARIMA model by evaluating multiple configurations using the AIC. To identify the optimal ARIMA model, different parameter combinations were tested, and their corresponding AIC values were recorded. The results are presented in Table 2.

Table 1. ADF test results

Variable	Value
Test Statistics	-3.21
Critical Value (5%)	-2.89
p-value	0.021
Decision	Stationary

Table 2. AIC values for different ARIMA models

Model	AIC value	Selected?
ARIMA (1,1,0)	320.5	No
ARIMA (2,1,1)	310.2	Yes
ARIMA (3,1,2)	315.7	No

Among the models tested, ARIMA (2,1,1) achieved the lowest AIC value (310.2), making it the optimal model for forecasting dengue cases in Pangasinan. This model was preferred as it effectively captured the temporal patterns in the dataset while avoiding unnecessary complexity. The selected model was then applied to generate forecasts, which are discussed in the following sections.

After finalizing the ARIMA (2,1,1) model, the next step was to evaluate its performance by testing it on the reserved dataset. The following plot compares the actual dengue cases with the forecasted values based on this train-test setup. This comparison provides a clear indication of the model's ability to predict future cases, demonstrating its effectiveness in forecasting dengue trends in Pangasinan.

Figure 3 presents the comparison between the forecasted and actual dengue cases in Pangasinan, illustrating the testing phase of the ARIMA model. As previously discussed in Tables 1 and 2, the model's training phase was numerically validated. To visualize the performance of the ARIMA (2,1,1) model, Figure 3 plots the historical and forecasted dengue case data. For model testing, 30% of the dataset, starting from mid-2022 through the remainder of 2024, was used, as indicated by the orange line in the figure. The forecasted dengue cases, generated using the trained ARIMA model, are shaded in dark blue. This forecast was derived by applying the model to the test dataset (30%), and the resulting predictions closely align with the historical data, demonstrating the ARIMA model's effectiveness in accurately predicting future dengue cases.

Although the model captures the overall trend, its predictive accuracy is limited, as it fails to precisely match the actual dengue case numbers. Despite these limitations, the model successfully identifies and projects the general seasonal pattern of dengue cases. A notable deviation occurs in 2024, where a sharp increase (orange line) in predicted cases stands out. This shift can be partially attributed to the reduced case numbers observed from 2019 to 2023. During the COVID-19 pandemic, people were less likely to seek hospital care for dengue symptoms [17], either due to fear of virus exposure [18] or restrictions on hospital admissions [19]. This led to an underreporting of dengue cases during this period, causing the data to reflect unusually low counts [20]. As a result, the model, which relies heavily on historical trends, may project a surge in cases post-pandemic due to the data discrepancies. This spike emphasizes the challenges of using historical data from irregular periods for accurate predictions. Additionally, environmental and climatic factors, such as rainfall and temperature, were not incorporated into the model. Integrating these variables in future studies may improve prediction precision [21]. Lastly, the model is more effective for short-term forecasting but may require recalibration if applied beyond a few years due to evolving disease transmission dynamics. Nevertheless, the model offers valuable insight into potential case trends, highlighting the need for cautious interpretation when analyzing pandemic-era data for future forecasting.

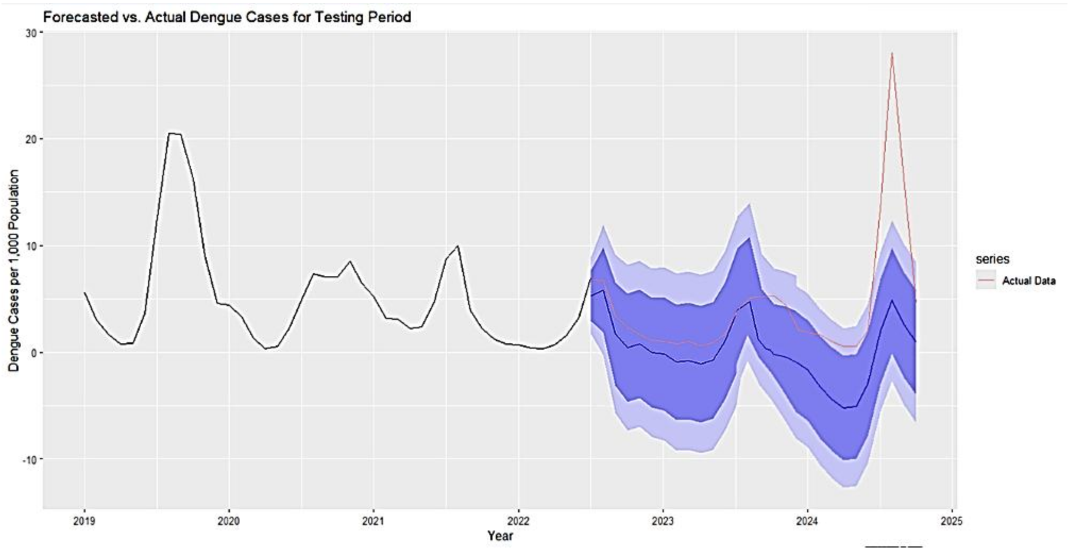

Figure 3. Comparison of actual and forecasted dengue cases in Pangasinan using ARIMA (2,1,1)

3.2.2. ARIMA model evaluation

To evaluate the accuracy of the ARIMA (2,1,1) model, the following statistical metrics were utilized: mean error (ME), RMSE, MAE, mean percentage error (MPE), MAPE, MASE, and the ACF. These metrics are widely accepted for assessing forecast performance, particularly for ARIMA models, as they measure prediction accuracy and offer insights into error characteristics [22], [23]. Table 3 presents the model's accuracy results based on these metrics.

Table 3. ARIMA model accuracy results and evaluation metrics

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-0. 0584633	1.142621	0.7401241	8.621976	21.69555	0.2733227	0.1425339

With a ME of -0.058 and a low MAE of 0.740, the model shows minimal bias and keeps average errors relatively small, suggesting reliable predictive performance. The RMSE of 1.143 indicates slightly larger errors are present but still within an acceptable range. A MAPE of 21.70% suggests moderate accuracy, with forecasted values deviating by an average of 21.70% from actual observations, which is generally acceptable in forecasting but could be improved. The MASE of 0.273 indicates that the model is significantly more accurate than a naive approach, while the low autocorrelation (ACF1 = 0.143) shows that residuals are mostly pattern-free, implying a good fit to the data's underlying structure.

To further evaluate the developed model, the MAPE is utilized as a key metric in calculating the accuracy percentage of the forecasted results. The percentage accuracy can then be derived as (3):

$$\begin{aligned} &Accuracy\ (\%) = 100 - MAPE \\ &Accuracy\ (\%) = 100 - 21.69555 \\ &Accuracy\ (\%) = 78.30445\% \end{aligned} \tag{3}$$

So, based on the MAPE of 21.70%, the computed accuracy suggests that, on average, the model's forecasts are accurate in capturing about 78.30% of the actual variation in dengue cases. This level of accuracy is generally acceptable, though the acceptable threshold often depends on the field and the criticality of predictions. Machine learning is subjective, and having a predictive accuracy above 70% is already considered a great model [24]. Kästner [25] claims that a model's accuracy depends entirely on the problem, and an accuracy range of 70%-90% is not only realistic but also considered useful and insightful. While the MAPE of 21.70% reflects some degree of error, it indicates that the model is still valuable for forecasting trends and providing actionable insights for public health decision-making. The findings of this study also align with previous research demonstrating the effectiveness of ARIMA in infectious disease forecasting. For instance, Othman *et al.* [8] applied ARIMA to dengue incidence in Surabaya and achieved similar seasonal forecasting accuracy, confirming ARIMA's robustness in capturing periodic trends. Likewise, several studies [26], [27] found that ARIMA performed comparably to LSTM for short-term infectious disease prediction, further validating the model's suitability for dengue forecasting in resource-limited settings. However, unlike national-scale models that generalize trends, this study focused on municipal-level predictions, allowing for targeted interventions at a granular level.

3.2.3. Forecasting dengue outbreaks in Pangasinan using ARIMA model

While the ARIMA model may not be flawless, it offers valuable insights into the general trends and patterns of dengue outbreaks in Pangasinan from 2019 to 2024. The close alignment between the observed and predicted cases in the testing period (as seen in Figure 3) highlights the model's effectiveness in capturing key temporal patterns. The model provides useful forecasts for future dengue cases, though it may not always capture sudden spikes or anomalies.

To further evaluate its utility, the ARIMA model was applied to forecast dengue cases in Pangasinan for the year 2025. This forecast builds on the historical trends observed in the previous years, offering projections that align with seasonal fluctuations in dengue incidence. Figure 4 presents the model's predicted dengue cases for 2025, highlighting the expected seasonal peaks and variations in case counts.

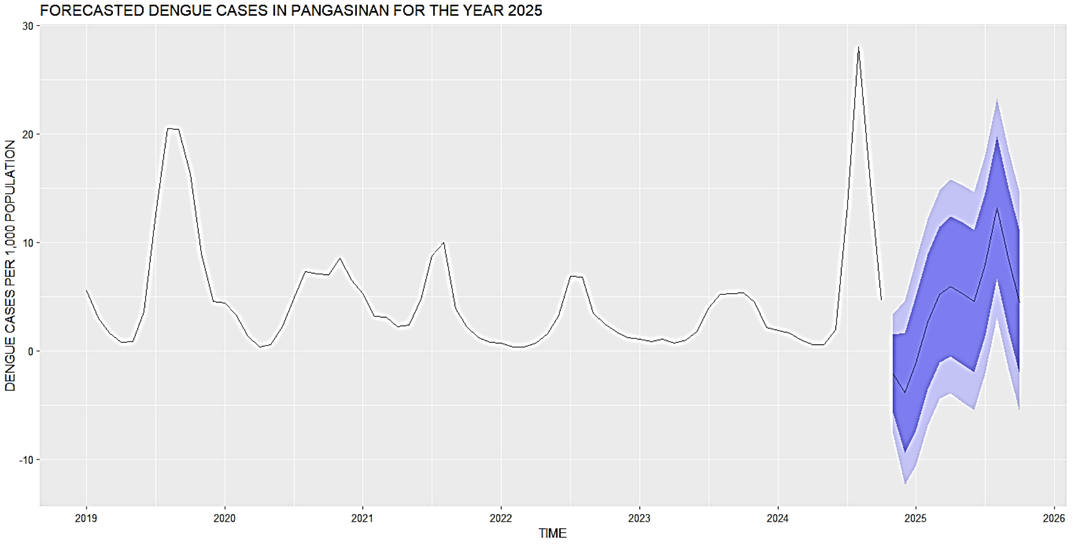

Figure 4. Forecasted dengue cases in Pangasinan for the year 2025

The ARIMA model's forecast for dengue cases in Pangasinan in 2025 reveals a clear anticipated rise, consistent with the seasonal patterns observed in previous years. As seen in the Figure 4, the forecasted dengue cases show a gradual upward trend beginning in the middle of the year and reaching a notable peak in the latter half of the year 2025. This pattern mirrors historical data, where dengue cases tend to surge during the rainy season when environmental conditions favor mosquito breeding. The forecasted trend highlights the model's ability to capture these recurring seasonal fluctuations, projecting an increase in cases as favorable conditions accumulate over time.

To further explore the model's performance in a more detailed time context, the researchers applied the ARIMA model to forecast monthly dengue cases for 2025. Building upon the forecasted rise in dengue cases for the latter half of 2025 (as shown in Figure 4), Figure 5 provides a more granular view by presenting the predicted monthly dengue cases in Pangasinan highlighting the exact periods that may experience the highest case counts. This detailed breakdown allows to pinpoint the specific months when the peak in dengue cases is expected to occur, offering a clearer picture of the anticipated seasonal fluctuations.

In Figure 5, the graph shows a substantial rise in dengue cases during the latter part of 2025, with a steady increase beginning in August and reaching a peak in October. The developed ARIMA model appears effective in capturing the seasonality and cyclical nature of dengue cases by projecting a peak period that aligns with observed historical trends. This suggests that the model accurately incorporates past seasonal fluctuations to anticipate future outbreaks. By accurately predicting the high-risk months, the model demonstrates its utility in guiding public health responses, allowing for targeted interventions in anticipation of peak dengue periods.

Furthermore, to assist local health units in their preparedness efforts, the ARIMA model was applied to forecast dengue cases at the municipal level. Figure 6 presents these forecasts, ranking municipalities based on projected dengue cases for 2025, from the highest to the lowest. This detailed breakdown not only reinforces the seasonal trends observed in the overall provincial forecast but also provides a more granular view, identifying which municipalities are most likely to experience the highest dengue outbreaks.

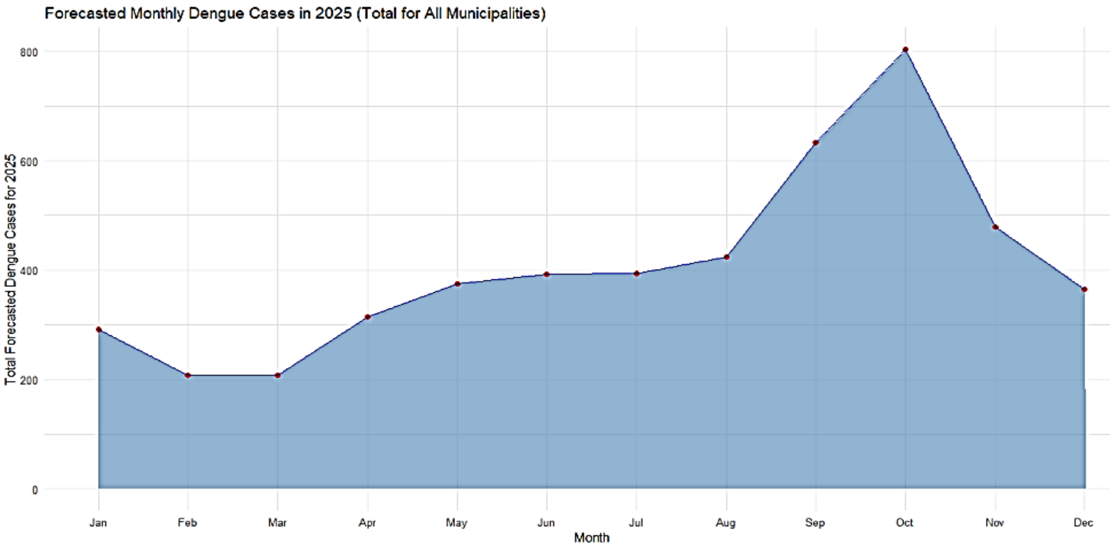

Figure 5. Forecasted monthly dengue cases for 2025 (Total for all municipalities)

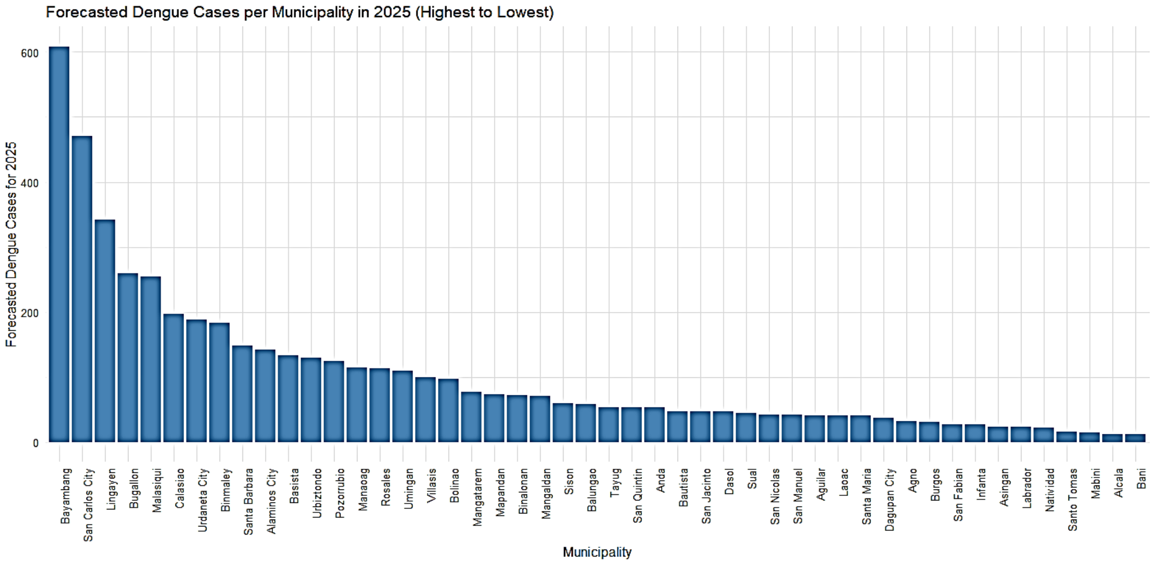

Figure 6. Forecasted dengue cases per municipality in 2025 (Highest to Lowest)

The ARIMA model appears to perform well in forecasting dengue cases, as its predictions for 2025 closely align with historical patterns observed from 2019 to 2024. The municipalities projected to have high dengue cases in 2025—such as Bayambang, Lingayen, San Carlos City, Malasiqui, and Bugallon—are consistent when compared with the areas that historically recorded higher cases, especially during peak months. Municipalities like Urdaneta City, Santa Barbara, and Alaminos City are in the moderate-risk category, while areas such as Agno, Burgos, and Infanta show low predicted case counts, reflecting historically lower incidence levels. This alignment suggests that the model effectively captures the underlying seasonal and geographical trends in dengue incidence, reinforcing its reliability for identifying high-risk municipalities.

4. CONCLUSION

This study successfully developed and validated an ARIMA model for forecasting dengue outbreaks in Pangasinan, utilizing historical data from 2019 to 2024. The model demonstrated reasonable predictive accuracy, achieving an average forecast accuracy of 78.3%, and effectively capturing the seasonal trends of dengue incidence in the region. The results highlighted potential increases in dengue cases for the year 2025, particularly during peak rainy months, aligning with observed historical patterns. Additionally, the forecasts identified high-risk municipalities, offering actionable insights for targeted public health interventions. These findings imply that the ARIMA model is a valuable tool for anticipating dengue outbreaks, providing a reliable basis for proactive planning and resource allocation. By enabling timely identification of high-risk periods and locations, this model can enhance public health responses, reduce the burden of dengue outbreaks, and contribute to more effective disease prevention strategies in Pangasinan.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Patrick Mole	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
Thelma Palaoag	✓	✓		✓					✓	✓	✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The dengue case data that support the findings of this study were obtained from the Provincial Epidemiology and Surveillance Unit of Pangasinan, Philippines. These data are available from the corresponding author upon reasonable request.

REFERENCES

[1] M. B. Khan, Z. S. Yang, C. Y. Lin, M. C. Hsu, A. N. Urbina, W. Assavalapsakul, and S. F. Wang, "Dengue overview: an updated systemic review," *Journal of Infection and Public Health*, vol. 16, no. 10, Aug. 2023, doi: 10.1016/j.jiph.2023.08.001.

[2] World Mosquito Program, "Dengue | World Mosquito Program," *Worldmosquitoprogram.org*, 2019. https://www.worldmosquitoprogram.org/en/learn/mosquito-borne-diseases/dengue

[3] S. Sangkaew, D. Ming, A. Boonyasiri, K. Honeyford, S. Kalayanarooj, and A. Holmes, "Risk predictors of progression to severe disease during the febrile phase of dengue: a systematic review and meta-analysis," *The Lancet Infectious Diseases*, vol. 21, no. 7, pp. 1014–1026, Jul. 2021, doi: 10.1016/s1473-3099(20)30601-0.

[4] V. Kaycee, "DOH worried as dengue cases rise by 68%," *RAPPLER*, Sep. 17, 2024. https://www.rappler.com/philippines/doh-report-dengue-cases-september-6-2024/

[5] "DOH: dengue cases lower in recent weeks - Department of Health," *Department of Health*, 2024. https://doh.gov.ph/press-release/doh-dengue-cases-lower-in-recent-weeks/

[6] H. Austria, "Dengue cases in Pangasinan posts 52% annual jump from Jan.1 to Aug. 5 | Philippine News Agency," *Pna.gov.ph*, 2024. https://www.pna.gov.ph/articles/1230765 (accessed Nov. 20, 2024).

[7] J. Kaur, K. S. Parmar, and S. Singh, "Autoregressive models in environmental forecasting time series: a theoretical and application review," *Environmental Science and Pollution Research*, vol. 30, no. 8, pp. 19617–19641, Jan. 2023, doi: 10.1007/s11356-023-25148-9.

[8] M. Othman, R. Indawati, A. A. Suleiman, M. B. Qomaruddin, and R. Sokkalingam, "Model forecasting development for dengue fever incidence in surabaya city using time series analysis," *Processes*, vol. 10, no. 11, p. 2454, Nov. 2022, doi: 10.3390/pr10112454.

[9] Y.-T. Tsan, D.-Y. Chen, P.-Y. Liu, E. Kristiani, L. Phuong, and C.-T. Yang, "The prediction of influenza-like illness and respiratory disease using LSTM and ARIMA," *International Journal of Environmental Research and Public Health/International Journal of Environmental Research and Public Health*, vol. 19, no. 3, pp. 1858–1858, Feb. 2022, doi: 10.3390/ijerph19031858.

[10] A. K. Sahai, N. Rath, V. Sood, and M. P. Singh, "ARIMA modelling and forecasting of COVID-19 in top five affected countries," *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, vol. 14, no. 5, pp. 1419-1427, Sep. 2020, doi: 10.1016/j.dsx.2020.07.042.

[11] X. Chen and P. Moraga, “Assessing dengue forecasting methods: A comparative study of statistical models and machine learning techniques in Rio de Janeiro, Brazil,” *medRxiv (Cold Spring Harbor Laboratory)*, Jun. 2024, doi: 10.1101/2024.06.12.24308827.

[12] M. Waqas and U. W. Humphries, “A critical review of RNN and LSTM variants in hydrological time series predictions,” *MethodsX*, vol. 13, pp. 102946–102946, Sep. 2024, doi: 10.1016/j.mex.2024.102946.

[13] F. Kamalov and H. Sulieman, “Time series signal recovery methods: comparative study,” *arXiv (Cornell University)*, Oct. 2021, doi: 10.1109/isncc52172.2021.9615669.

[14] Q. H. Nguyen, H. B. Ly, L. S. Ho, N. Al-Ansari, H. V. Le, V. Q. Tran, and B. T. Pham, “Influence of data splitting on performance of machine learning models in prediction of shear strength of soil,” *Mathematical Problems in Engineering*, vol. 2021, pp. 1–15, Feb. 2021, doi: 10.1155/2021/4832864.

[15] B. Vrigazova, “The proportion for splitting data into training and test set for the bootstrap in classification problems,” *Business Systems Research Journal*, vol. 12, no. 1, pp. 228–242, May 2021, doi: 10.2478/bsrj-2021-0015.

[16] A. L. Saipen, B. Demot, and L. De Leon, “Dengue–COVID-19 coinfection: the first reported case in the Philippines,” *Western Pacific Surveillance and Response Journal*, vol. 12, no. 1, pp. 35–39, Mar. 2021, doi: 10.5365/wpsar.2020.11.3.016.

[17] D. N. Mashudi, N. Ahmad, and S. M. Said, “Level of dengue preventive practices and associated factors in a Malaysian residential area during the COVID-19 pandemic: a cross-sectional study,” *PLOS ONE*, vol. 17, no. 4, p. e0267899, Apr. 2022, doi: 10.1371/journal.pone.0267899.

[18] R. Moore, R. S. Purvis, E. Hallgren, S. Reece, A. Padilla-Ramos, M. Gurel-Headley, and P. A. McElfish, “‘I am hesitant to visit the doctor unless absolutely necessary’: A qualitative study of delayed care, avoidance of care, and telehealth experiences during the COVID-19 pandemic,” *Medicine*, vol. 101, no. 32, p. e29439, Aug. 2022.

[19] M. Muselli *et al.*, “The impact of COVID-19 pandemic on emergency services,” *Annali Di Igiene: Medicina Preventiva E Di Comunita*, vol. 34, no. 3, pp. 248–258, 2022, doi: 10.7416/ai.2021.2480.

[20] H. Sharma, A. Ilyas, A. Chowdhury, N. K. Poddar, A. A. Chaudhary, S. A. R. Shilbayeh, and S. Khan, “Does COVID-19 lockdowns have impacted on global dengue burden? A special focus to India,” *BMC Public Health*, vol. 22, no. 1, Jul. 2022, doi: 10.1186/s12889-022-13720-w.

[21] X. Y. Leung, R. M. Islam, M. Adhami, D. Ilic, L. McDonald, S. Palawaththa, and M. N. Karim, “A systematic review of dengue outbreak prediction models: Current scenario and future directions,” *PLOS Neglected Tropical Diseases*, vol. 17, no. 2, p. e0010631, Feb. 2023, doi: 10.1371/journal.pntd.0010631.

[22] K. D. Ligue and K. J. Ligue, “Deep learning approach to forecasting dengue cases in davao city using long short-term memory (LSTM),” *Philippine Journal of Science*, vol. 151, no. 3, Mar. 2022, doi: 10.56899/151.03.01.

[23] G. W. Khamala, J. W. Makokha, and R. Boiyo, “Statistical analysis of aerosols characteristics from satellite measurements over east africa using autoregressive moving average (ARIMA),” *OALib*, vol. 09, no. 11, pp. 1–14, 2022, doi: 10.4236/oalib.1109496.

[24] K. Barkved, “How to know if your machine learning model has good performance | obviously AI,” *www.obviously.ai*, Mar. 09, 2022. https://www.obviously.ai/post/machine-learning-model-performance

[25] C. Kästner, “Model quality: measuring prediction accuracy,” *Medium*, Mar. 22, 2021. https://ckaestne.medium.com/model-quality-measuring-prediction-accuracy-38826216ebcb

[26] S. M. Khan, “Application of deep learning LSTM and ARIMA models in time series forecasting: a methods case study analyzing canadian and swedish indoor air pollution data,” *Austin Journal of Medical Oncology*, vol. 9, no. 1, Dec. 2022, doi: 10.26420/austinjmedoncol.2022.1073.

[27] S. Sah, B. Surendiran, R. Dhanalakshmi, S. N. Mohanty, F. Alenezi, and K. Polat, “Forecasting COVID-19 pandemic using prophet, ARIMA, and hybrid stacked LSTM-GRU models in India,” *Computational and Mathematical Methods in Medicine*, vol. 2022, pp. 1–19, May 2022, doi: 10.1155/2022/1556025.

BIOGRAPHIES OF AUTHORS

Patrick Mole earned his master’s degree in Information Technology (MIT) from Saint Louis University in Baguio City, Philippines, in 2022, where his major field of study was research. He is currently pursuing his Doctor of Information Technology (DIT) at the University of the Cordilleras. His research interests include data analytics and machine learning. Mr. Mole is currently an IT instructor at Urdaneta City University in Pangasinan, Philippines, and is a member of the Philippine Society of Information Technology Educators (PSITE). He can be contacted at email: pvm8598@students.uc-bcf.edu.ph or patrickmole@ucu.edu.ph.

Thelma Palaoag is a director of the Innovation and Technology Transfer Office of the University of the Cordilleras. She is a visionary leader, accomplished researcher, and trailblazer in the field of Information Technology (IT). With an unwavering commitment to advancing technological frontiers, she has dedicated her career to pushing the boundaries of IT research and fostering innovation within academic institutions. She can be contacted at email: tdpalaoag@uc-bcf.edu.ph.