

Autoregressive integrated moving average-long short-term memory optimized hybrid model for cybercrime forecasting

Manuel Martin Morales-Barrenechea^{1,2}, Ciro Rodriguez¹, Ernesto David Cancho-Rodriguez¹, Ricardo Richard Huamantingo Navarro¹

¹Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional Mayor de San Marcos, Lima, Perú

²Facultad de Ingeniería de Redes y Comunicaciones, Universidad Peruana de Ciencias Aplicadas, Lima, Perú

Article Info

Article history:

Received Dec 10, 2024

Revised Sep 3, 2025

Accepted Sep 11, 2025

Keywords:

Autoregressive integrated
moving average
Cybercrime
Hybrid
Long short-term memory
Optimization

ABSTRACT

Cybercrime represents a growing global threat with adverse impacts on citizen security, the digital economy, and quality of life. In this context, an optimized hybrid model was developed that combines autoregressive integrated moving average (ARIMA) and long short-term memory (LSTM) for the monthly forecast of cybercrime complaints, applying the cross industry standard process for data mining (CRISP-DM) methodology and applying Python based data science techniques. The model combines the capabilities of the ARIMA statistical approach to capture linear components with the power of LSTM neural networks to address nonlinear temporal relationships. The architecture was trained on a set of 60,378 official records of complaints registered by the National Police of Peru between 2018 and 2023, achieving a mean absolute percentage error (MAPE) of 10.73%, which represents a significant improvement over the singular ARIMA and LSTM predictive models. Compared to previous studies in crime, health, and agriculture, this approach showed a greater ability to generalize over complex time series. It is concluded that the application of the proposed model is a relevant contribution for the police and other security agencies to anticipate crime trends and design preventive and effective strategies to combating cybercrime.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Manuel Martin Morales-Barrenechea
Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional Mayor de San Marcos
Av. Carlos Germán Amezaga 375, Lima, Perú
Email: martin.moralesb@unmsm.edu.pe

1. INTRODUCTION

Cybercrime has become a relevant world threat [1], [2]. It is a problem that undermines people's standard of living [3] and the economic development of countries [4]. Cybercrime not only affects the security of people's and organizations' data [2], [5], but also paralyzes commerce, generating a loss of productivity [6]. Technological progress has intensified society's use of digital mediums for exchanging information and facilitating commerce [7], and the COVID-19 pandemic, cybercrime led to a marked escalation in incidents in the crime rate [1], [2], [8]. The cost generated by cybercrime reached \$8 trillion in 2023 and is estimated to reach \$10.5 trillion by 2025 [7], [9].

The use of artificial intelligence to combat cybercrime is increasing. Researchers are increasingly employing machine learning techniques to analyse and predict cybercrime patterns [10], achieving high rates of accuracy in detecting and predicting crimes [11]. Traditional techniques have laid the groundwork for predicting cybercrime [4]. Traditional models such as decision tree, random forest, and support vector machine are commonly utilized for predicting criminal activity, given their capability to analyse organized

datasets and provide interpretable results [4], [12]. However, traditional models often struggle in the face of changing cyber threat dynamics [13]. Therefore, the authors suggest the use of hybrid models to address these limitations [14], [15].

The integration of the autoregressive integrated moving average (ARIMA) and long short-term memory (LSTM) models improves forecast accuracy by leveraging their strengths in handling linear and nonlinear trends in time series data [16], [17]. The ARIMA model effectively detects linear trends and seasonality [16], [18], while the LSTM excels at modelling complex nonlinear relationships and long-term dependencies [17], [19]. The integration of both models allows for a more thorough analysis of time series data, addressing the limitations of each method when used independently. The ARIMA–LSTM hybrid approach is generally recognized as a promising strategy for improving prediction accuracy in various fields [18], [20] and providing insights into crime patterns and trends [16], [17].

The ARIMA-LSTM hybrid model has recently been used as a superior option in many forecasting tasks due to its ability to reduce error and variance compared to individual models [17], [19] that have been applied in various sectors. In the agriculture sector, an ARIMA-LSTM model was applied for the prediction of shallot prices in Thailand [17], with a mean absolute percentage error (MAPE) of 13.6% indicating that the combined model achieved better results than traditional machine learning models [15]. In the construction industry [19] addressed the forecasting of demand in the sector through the hybrid model, with which a MAPE of 13.3% was achieved, notably lower in comparison to the results of the separate models.

In the health sector, Deng *et al.* [14] presents a hybrid ARIMA-LSTM model tuned to predict outpatient demand with lower results and a MAPE of 14.1% compared to individual models, which indicates more accurate predictions. In information technology, Liu *et al.* [20] developed the linear state-space autoregressive (LSAR) model that combines the ARIMA and LSTM models to accurately predict cloud computing resource loads with a MAPE of 3.81%. In meteorology, Xu *et al.* [15] proposed a hybrid approach designed to enhance the precision of short-term drought forecasting in China with a MSE of 0.347, and the study [21] developed a hybrid model to predict meteorological variables and anticipate anomalies, such as strong winds and snowstorms, reaching an accuracy of 95% to anticipate these anomalies.

On the crime front, the hybrid ARIMA-LSTM model has been explored in crime prediction in India [10] which demonstrated that the combination of ARIMA-LSTM resulted in more accurate crime rate predictions compared to traditional models. This combination of models proved effective in predicting crime rates influenced by factors such as socioeconomic conditions and law enforcement practices [10], [18]. Another study [22] the hybrid model was applied to forecast crime rates in urban settings, which demonstrated superior performance compared to traditional models with root mean square error (RMSE).

However, considering this research, we find gaps in the use of the ARIMA–LSTM hybrid framework, especially in the field of cybercrime prediction, and limitations in the optimization of the hybrid architecture to improve forecast prediction. Therefore, this research aims to build an optimized hybrid ARIMA–LSTM model to reduce error in predicting cybercrime forecasting as shown in Figure 1. Finally, this optimized hybrid model aims to be a methodological contribution to the knowledge of predictive analysis and represents a promising alternative to advance in the field of forecasting cybercrime complaints, which contributes to the police and authorities to formulate strategies and prevention measures in the fight against cybercrime.

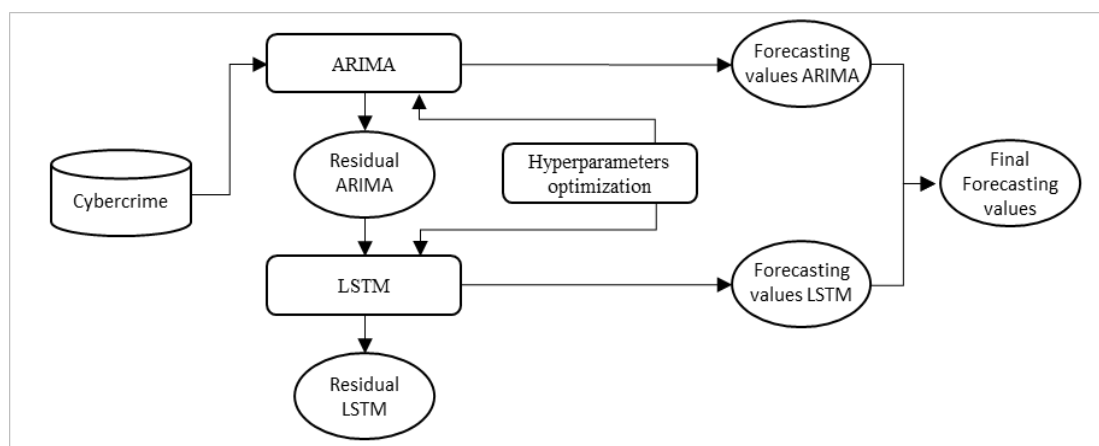


Figure 1. ARIMA-LSTM optimized model

2. METHOD

The cross industry standard process for data mining (CRISP-DM) methodology was used as a procedure to develop the proposed model [23], [24] in the six phases proposed by the procedure and described in the following sections. This methodology used for data mining projects has also been applied to develop machine learning-based classification models to detect cybercrime [25]. For the construction of the model, the Python programming language was used through the Jupyter Notebook application.

2.1. Understanding the environment

This research aims to develop an optimized combined ARIMA and LSTM models to reduce error in the prediction of cybercrime reporting forecasting. They correspond to the cases reported to the National Police of Peru within the framework of law No. 30096 - law on computer crimes. Its practical application is relevant for police and other institutions to anticipate cybercrime behaviour, establish strategies, and strengthen the fight against cybercrime.

2.2. Understanding the data

The dataset used refers to the complaints registered in the police complaints system (SIDPOL) between January 2018 and December 2023. The data present values related to the time of the occurrence of the cybercrime, the gender of the cybercriminal, the type of cybercrime, and the place where the event was reported. These reports of cybercrime are equivalent to a total of 75,471, of which 80% (60,386) are for computer fraud and those used in this study.

2.3. Data pre-processing

The cybercrime dataset should be pre-processed appropriately before being used as model inputs. It is necessary to preprocess the original time series to validate and obtain better training results of the model [15]. A stationary time series is a prerequisite for the reliable use of the ARIMA forecasting method [15]. Normalized data eliminates the impact on the neural network and facilitates analysis by improving the speed of model training. Noise and outliers are a common phenomenon in crime data [22]. Data preprocessing strengthens data planning and contributes to generating statistically significant data to make accurate predictions about the rate of cybercrime [26].

The flowchart for data preprocessing is shown in Figure 2. First, data cleansing removes records outside the analysis time range and duplicate records. Consistency testing is then performed to identify and remove records that have empty or null values. Next, the time series of data is structured based on the periodic grouping of time. Then, to confirm stationarity in the time series, the augmented dickey-fuller (ADF) test is applied to validate the stationarity if the p-value is under the 0.05 threshold. If it is not stationary, the differentiation of the series is applied until the new transformed or stationary series is achieved, which is a key requirement for the ARIMA model.

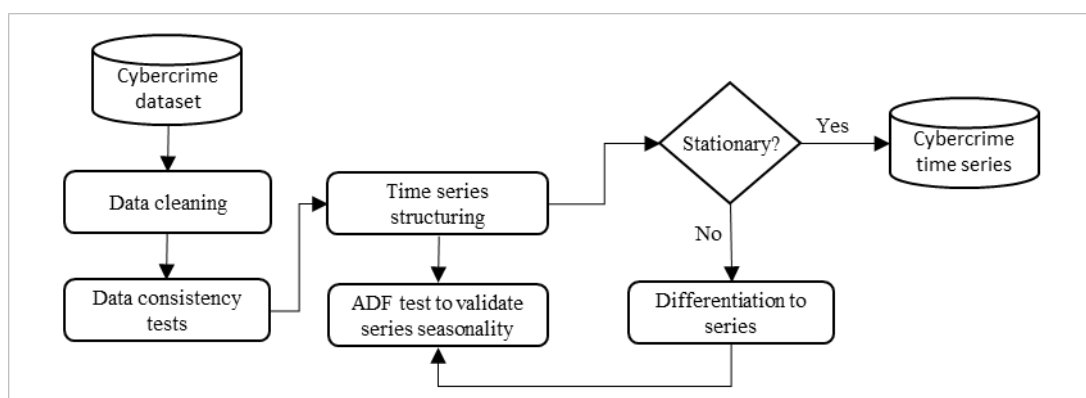


Figure 2. Data preprocessing flowchart

2.4. Modelling

2.4.1. Autoregressive integrated moving average

This model offers a statistical basis for analysing trends and generating forecasts from time series data [20] and is also known as the Box Jenkins model [27]. The ARIMA model, denoted by (p, d, q) , integrates two components: the autoregressive (AR) part and the moving average (MA) part. In this context,

p signifies the number of lagged observations in the AR component, d represents the number of times the data needs to be differenced to become stationary, and q refers to the number of lagged forecast errors included in the MA part [20]. The following equation shows the general structure of the ARIMA model:

$$L_t = \sum_{i=1}^p \phi_i L_{t-i} + \varepsilon_t - \sum_{j=1}^q \theta_j \varepsilon_{t-j}$$

where L_t is the cluster load value at time t , ϕ , and θ are the AR parameter and the MA parameter respectively; p is the order of the AR model; q is the order of the MA model; and ε is the residual sequence.

The modelling of the ARIMA model follows the steps: i) verify the stationarity of the time series, if it is not stationary, differentiation is performed using the ADF test where d is determined; ii) identify p and q using the autocorrelation function (ACF) and partial autocorrelation function (PACF) graphs; iii) comparing Akaike information criterion (AIC) and Bayes information criterion (BIC); and iv) the ARIMA parameter estimation. This research proposes and considers in the preprocessing of the data, the verification of seasonality and the differentiation of the time series.

2.4.2. Long short-term memory network

Based on the original study by Graves and Schmidhuber, LSTM networks are a specialized type of recurrent neural network (RNN) Hernández *et al.* [28]. The purpose of the model is to retain information for extended periods, allowing for the identification of long-term patterns within time series trends. It consists of memory cells and gate mechanisms namely the input gate i_t , the forget gate f_t , and the output gate o_t . These gates manage the information flow into, within, and out of the memory cell. At distinct time intervals, the cell stores values, and these gates selectively allow data passage based on computed weights [20].

The activation function σ maps input to the range (0,1), and \odot denotes element-wise multiplication. c_t refers to the cell state, and x_t is the current input vector. The vectors h_{t-1} and h_t represent the hidden states at times $t-1$ and t , respectively. The parameters W , U , and B are learnable during training. The LSTM's internal operations can be described step-by-step as follows [20]:

Gate equations:

$$i_t = \sigma(Wx_t + U_i h_{t-1} + B_i)$$

$$f_t = \sigma(Wx_t + U_f h_{t-1} + B_f)$$

$$o_t = \sigma(Wx_t + U_o h_{t-1} + B_o)$$

Cell state update:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(Wx_t + U_c h_{t-1} + B_c)$$

Hidden state output:

$$h_t = o_t \odot \tanh(c_t)$$

These formulas reflect how prior memory c_{t-1} and current input x_t interact to yield updated memory and hidden states. The computation occurs in three stages: i) state generation from past hidden state and input, ii) update of the cell memory, and iii) derivation of the current hidden state using the output gate.

2.4.3. Autoregressive integrated moving average-long short-term memory optimized hybrid model

The research implemented an optimized ARIMA-LSTM for the prediction of cybercrimes due to computer fraud [29]. According to Kasemset *et al.* [17], ARIMA-LSTM hybrid has proven effective in delivering notable performance gains forecast accuracy by combining the strengths of both approaches in time series analysis. The combination of these models involved selecting and optimizing the parameters, following the assessment of seasonality for ARIMA modelling, the LSTM model was trained using the resulting residual errors obtained from ARIMA to capture the nonlinear and unexplained components by the linear model [16], [18], improving the accuracy of the forecast by a hybrid architecture [15] represented by the following formula:

$$Y_t = L_t + N_t$$

where Y_t is composed of linear (L_t) and nonlinear (N_t) components.

The architecture of the proposed solution is shown in the flow chart of the Figure 3. First, the preprocessing of the dataset is considered in a sequence of steps that are detailed in the preprocessing section. The ARIMA model is first utilized to analyze and represent the linear trends. Following the ARIMA modelling, its residual errors are fed into the LSTM model to handle the remaining nonlinear trends in the time series. In the development of the models, hyperparameter optimization techniques were applied in both models to improve their predictive performance. In the last step, the optimized hybrid model ARIMA-LSTM generates the final forecast values for evaluation.

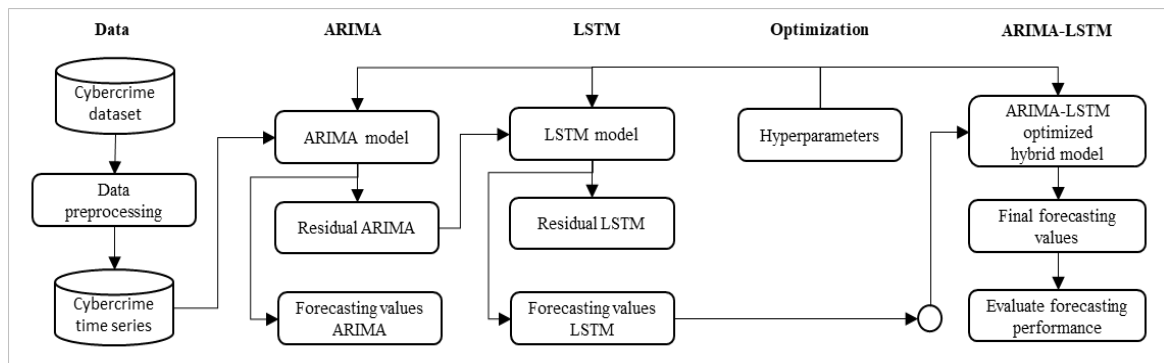


Figure 3. ARIMA-LSTM optimized hybrid model flowchart

2.5. Evaluation of results

The prediction accuracy of each forecast model was assessed with the RMSE, mean absolute error (MAE), and MAPE indicators. A decrease in these evaluation criteria suggests an improvement in the predictive capabilities of the ARIMA-LSTM model. The formulas used to calculate these criteria [17], [20], [30], [31] are shown in Table 1.

Table 1. Equations used for model evaluation

Criteria	Evaluation	Formula	Values
RMSE	Square root of the MSE and provides a measure of error in the same units as the original data	$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$	A 0 value indicates that the expected and actual values match precisely. Smaller RMSE scores are associated with higher prediction accuracy, while larger scores signal greater forecasting errors.
MAE	Measure the average magnitude of errors in a set of predictions, regardless of their direction	$MAE = \frac{1}{n} \sum_{i=1}^n y_i - \hat{y}_i $	A 0 value would mean a perfect prediction. The smaller the MAE, the better the model's predictions.
MAPE	Percentage measure of prediction accuracy, which is useful for understanding the error relative to actual values	$MAPE = \frac{1}{n} \sum_{i=1}^n \left \frac{y_i - \hat{y}_i}{y_i} \right $	Highly accurate forecasting <10% Good forecasting 10%-20% Reasonable forecasting 20%-50% Weak an inaccurate forecasting >50%

In these formulations, y_i denotes the true observed value, \hat{y}_i refers to the estimated value produced by the forecasting model, and n indicates the total number of forecasted time steps. The low values of these indicators demonstrated a slight variation between the actual and forecasted data. The model with the lowest criterion values obtained is chosen as the most appropriate model [19], [32].

2.6. Deployment

The results of the ARIMA-LSTM optimized hybrid model have relevant practical implications. Institutions such as the National Police of Peru and the Public Prosecutor's Office can use the model to anticipate future behaviours in the reporting of computer crimes, allowing a more efficient planning of resources, design of preventive campaigns and targeting of operational interventions. The predictive value of the model could help move from a reactive to a proactive logic in the fight against cybercrime.

3. RESULTS AND DISCUSSION

3.1. Data processing

As part of the consistency verification process, 8 records that were outside the 2018 and 2023 range were eliminated, leaving a total of 60,378 records. No duplicate records were found and the fields did not contain empty or null values. Then, the time series was structured based on the monthly grouping of the date of the complaint and quantity, where 72 series were obtained from January 2018 to December 2023. Next, the ADF test was performed to confirm the presence of stationarity in the time series, a p-value of $0.991 > 0.05$ was obtained, so the series was considered non-stationary in its original form, and that was confirmed with the ACF and PACF graphs of Figure 4. Then, differentiation was applied to the *series.diff()* finally obtains a transformed or stationary series (Figure 5), which is a key requirement for the ARIMA model.

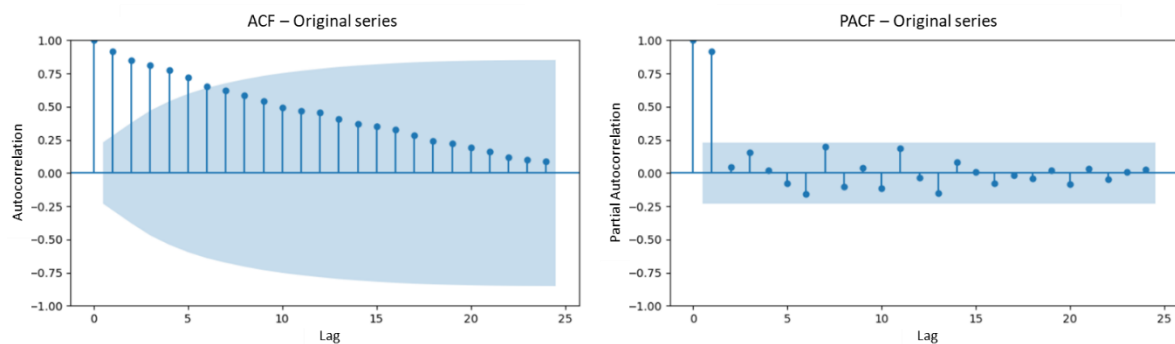


Figure 4. ACF and PACF chart of the original series (non-stationary)

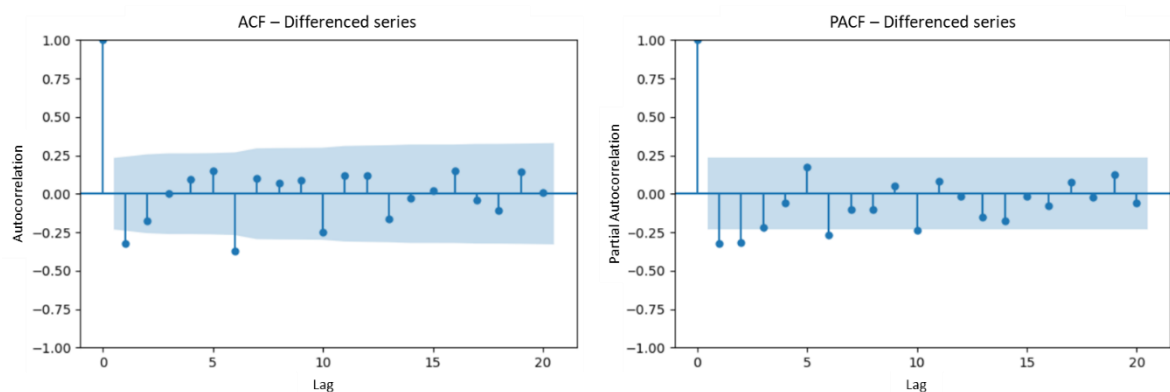


Figure 5. ACF and PACF chart of the differenced series (stationary)

3.2. Autoregressive integrated moving average modelling

For the development of the ARIMA model, p parameters were estimated using the ACF plots and the potential q parameters were inferred by analysing the PACF plots. To determine the optimal model, the AIC and BIC criteria were selected. Finally, the optimal ARIMA (1,1,2) was selected with an AIC of 916.58. The performance results of the optimized ARIMA model were for MAE of 104.39, a RMSE of 143.95, and a MAPE of 15.43.

Figure 6 shows the performance of the ARIMA forecast. The blue trajectory visualizes the original time series of monthly complaints between 2018 and 2023, while the orange line shows the prediction generated by the ARIMA model during the training and testing period. A good fit of the model to the historical data is observed, capturing the growing trend. The green line shows the future projection for the first half of 2024, where the behaviour suggests a possible stabilization of cybercrimes.

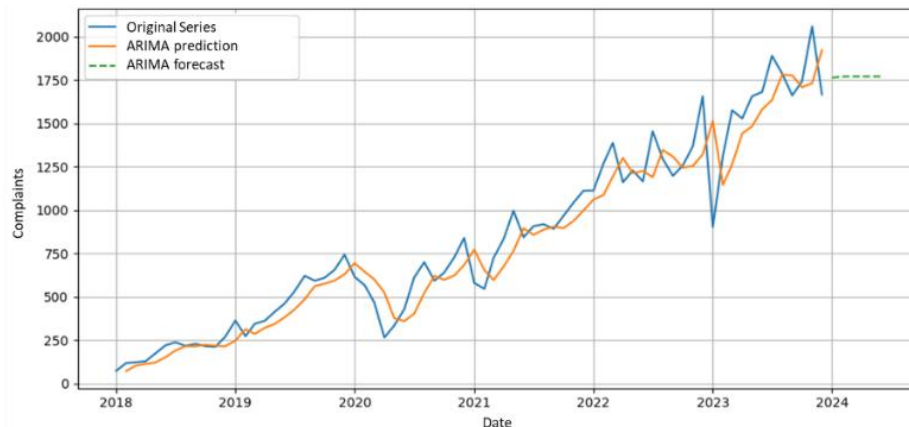


Figure 6. Comparison of original series, prediction and forecast of ARIMA model

3.3. Long short-term memory modelling

For the development of the LSTM model, the MinMaxScaler technique helped normalize the input distribution and streamline the training process of the model [23], [24]. Six sequence steps or previous observations were used as inputs to estimate the subsequent value in the temporal sequence, allowing the capture of short-term temporal dependencies. The division of the time series was performed assigning 60% to training and 40% to the testing phase.

The model underwent training over 50 epochs, which allowed for a progressive adjustment of the network weights. Hyperparameter optimization was performed using the random search technique, evaluating multiple combinations of parameters such as the number of LSTM units, activation functions, and optimization algorithms. This process was complemented with an internal cross-validation scheme and the use of Early Stopping to stop training early in case of overfitting.

Figure 7 shows the effectiveness of the LSTM architecture. The blue line illustrates the observed time series, while the orange line indicates the projection generated by the model during the validation period. The model manages to capture the general trend, although with less sensitivity to abrupt fluctuations. The green line corresponds to the future projection for the first half of 2024. The projected trajectory shows a slowdown in the pace of growth, which could reflect an eventual slowdown in cybercrime.

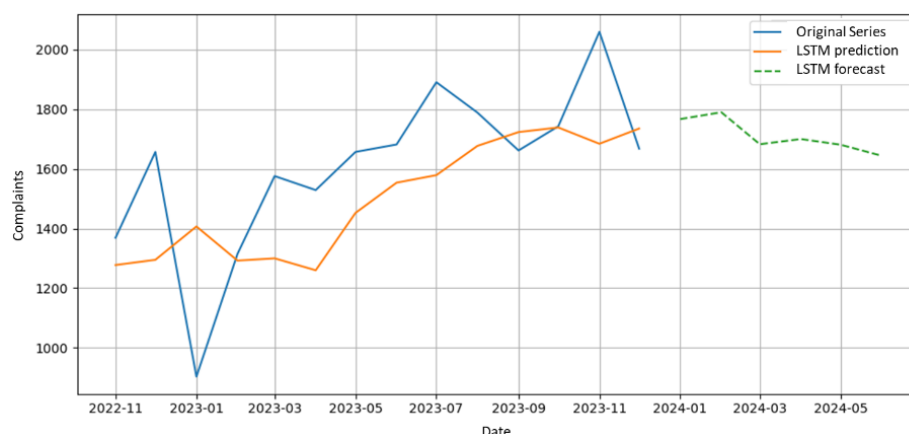


Figure 7. Comparison of original series, prediction, and forecast of LSTM model

3.4. Autoregressive integrated moving average-long short-term memory optimized modelling

The optimization of the combined ARIMA-LSTM architecture through the configuration of the hyperparameters was key to improving the predictive capacity of the model. Techniques were used to automate the process of selecting the combinations of the parameters with some manual adjustments to achieve optimal regression performance. The selection techniques and hyperparameters used in the regression models are shown in Table 2.

Table 2. Hyperparameter setting

Regression	Hyperparameter
ARIMA	<ul style="list-style-type: none"> - Technique: <i>auto ARIMA</i> to obtain the parameters p, d, and q of the model ARIMA. - Hyperparameters: seasonal=false, suppress warnings=true, max p=4, max q=4, trace=true, and scoring=MSE.
LSTM	<ul style="list-style-type: none"> - Technique: <i>random search</i> to randomly select combinations of hyperparameters (hidden units, activation function, and optimizer). - Hyperparameters: LSTM units=50, dropout rate=0.25, activation='ReLU', epochs=200, optimizer='Adam', batch size=12, and loss function=MSE.

In the construction of the ARIMA-LSTM combined architecture, the model ARIMA (2,1,2) was selected because of the adjustment process. Subsequently, the Ljung-Box test was applied, obtaining a p-value of 0.77, higher than the significance threshold of 5%, which indicated that the residuals of the model do not present significant autocorrelation and can be considered white noise. This feature allowed them to serve as input to the LSTM based model, responsible for capturing the remaining nonlinear patterns.

Each of the models, ARIMA and LSTM were optimized using hyperparameter tuning techniques described in Table 2. For the LSTM model, the temporal series division was performed 60-40 for training and test data, respectively. As a result of the optimization process, the ARIMA-LSTM hybrid model achieved outstanding performance, achieving a MAE of 150.23, a RMSE of 195.84, and a MAPE of 10.73%, a value that, according to the criteria established in Table 3. Comparison of criteria between ARIMA, LSTM, and ARIMA-LSTM optimized model, is classified as highly accurate predictive.

The Figure 8 demonstrates the effectiveness of the ARIMA-LSTM optimized. The blue line corresponds to the original series of monthly complaints between 2018 and 2023, with some seasonality and variability since 2020 that reflects the sustained growth of cybercrimes, while the orange line represents the adjustment of the hybrid model on historical data. It is observed that the model manages to capture the growing trend and part of the short-term variability. The dotted green line shows the projection for the first half of 2024. The correlation between the original series and the model's prediction demonstrates an effective capture of the temporal patterns validated by a MAPE of 10.73%.

Table 3. Comparison of criteria between ARIMA, LSTM and ARIMA-LSTM optimized model

Model	Criteria		
	MAE	RMSE	MAPE (%)
ARIMA	104.39	143.95	15.43
LSTM	196.75	249.25	13.28
ARIMA-LSTM optimized	141.31	188.64	10.73

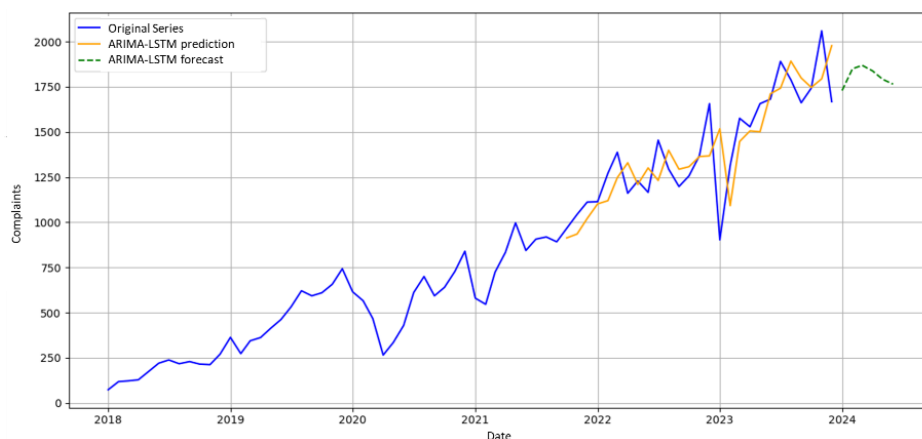


Figure 8. Comparison of original series, prediction, and forecast of ARIMA-LSTM optimized model

In summary, the optimized fusion of ARIMA and LSTM models demonstrates a good ability to accurately model historical patterns of the series to provide a realistic and useful forecast for decision making. Its ability to model time series with high variability makes it a valuable tool for the early detection of crime trends for the strategic planning of police and other institutions in the prevention and response to cybercrime.

3.5. Discussion

This study developed and validated an optimized the ARIMA-LSTM integrated framework that merges the strengths of ARIMA, which handles linear aspects, and LSTM, which addresses nonlinear components [14], [17] to enhance the precision of predictions related to cybercrime reports. Although earlier research explored the influence of conventional machine learning approaches that laid the foundation for cybercrime prediction [11], these individual models presented limitations [18], [20] where the hybrid approach is presented as a promising strategy to improve prediction accuracy in various fields [16], [17].

It was found that while the ARIMA-LSTM hybrid model demonstrates superior performance compared to traditional models in terms of performance metrics such as RMSE, MAE, and MAPE [14] in forecast accuracy. In addition, it is advantageous in scenarios where individual models do not capture all patterns present in the data [17] and where opportunities for improvement were found through hybrid model optimization. Therefore, the combined ARIMA-LSTM was optimized through hyperparameter configuration and neural network backpropagation to obtain higher prediction accuracy compared to independent models [14].

Among other findings, we found that the timescale of the data influences the performance of hybrid models such as ARIMA-LSTM, which improves with increasing time [15]. The study [10] used historical data from 9,840 records on crimes in India between 2001 and 2013. Sharmin *et al.* [22] used data on London crime from 2014 to 2020 while this study used a dataset that considered 60,378 complaints between 2018 and 2023. Not only is the use of historical data [19] important for the adjustment of the hybrid model, but also for the processing [22], relevance and reliability [19] of the data for analysis. In addition, the division of datasets into training and testing models contributes to model accuracy and reliability [14].

Our results show that the optimized combination of ARIMA-LSTM improves MAPE (10.73%) by 19% in comparison with the individual ARIMA (15.50%) and LSTM (13.70%) architectures in the prediction of cybercrimes. These results align with [22] of London crime prediction model that also outperformed individual models, reaching an accuracy of 97% against ARIMA (89%) and the neural network model (87%). Unlike this study, which applied the Auto Arima function for parameter selection and a stacked architecture for LSTM, in this study, we opted for hyperparameter-based optimization in both models.

We also compared this proposal with other applied studies and discovered that in the health sector [14] a hybrid model was applied to a dataset of medical visits with a weekly frequency and a time window of 24 weeks, with a MAPE of 14.13% and in the agriculture sector [17] used a time series of 84 months on agricultural prices where the ARIMA-LSTM model obtained a MAPE of 13.62%. However, neither study exceeded the performance achieved in this study with an MAPE of 10.73% trained on a more extensive basis of 72 months, which indicates a greater generalizability and robustness in the face of complex temporal patterns.

Our findings provide definitive evidence that the hybrid model obtains a better MAPE of 10.73% in comparison with the individual ARIMA (15.50%) and LSTM (13.70%) architectures. Undoubtedly, improvement related to a complementary architecture approach of both models and the optimization of hyperparameters. In addition, compared to other studies from different sectors, the hybrid model optimized for cybercrime forecasting stands out with a lower MAPE compared to other studies applied in different fields such as agriculture, construction, health, among others.

Finally, our optimized hybrid model ARIMA-LSTM allowed us to capture both linear trends and nonlinear variations more accurately. This integration capacity has resulted in a better relative and absolute performance in the forecasting of cybercrime complaints. Although the findings of this study reaffirm the usefulness of the hybrid approach, we recognize some limitations, such as the incorporation of exogenous variables, such as technological events or public policies, that could require additional research to confirm their contribution to the predictive capacity of the model. Future research could analyse and replicate its application to other social or criminal problems by considering external factors to confirm and expand our findings.

4. CONCLUSION

This research developed and validated an optimized hybrid model, ARIMA-LSTM, for the monthly forecast of complaints of computer crimes against property in Peru, effectively integrating linear and nonlinear components in a complementary architecture. The model was built under the CRISP-DM methodology, implemented in Python, and trained on a set of 60,378 official records from the National Police of Peru, processed as a time series of 72 months. By optimizing hyperparameters by auto ARIMA, random search, and early stopping, the model achieved a MAPE of 10.73%, statistically outperforming the individual models. Compared to previous studies in various areas, such as crime, health, and agriculture, the proposed model showed a greater capacity for generalization and adaptation to complex temporal patterns. From an applied perspective, the model offers the National Police of Peru and other entities a tool to anticipate crime

patterns and design strategies in the fight against cybercrime. Finally, although the effectiveness of the hybrid approach has been proven, the exclusion of exogenous variables related to public policies, social dynamics, or technological events is recognized as a limitation. Therefore, future research may benefit from the inclusion of these factors, as well as the applicability of the model in other criminal or social contexts, to improve the predictive capacity of the model.

FUNDING INFORMATION

The authors express their gratitude to the Universidad Nacional Mayor de San Marcos for its institutional support. This study is part of the lead author's doctoral research and was funded equally by all authors.

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
Manuel Martin	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Morales-Barrenechea														
Ciro Rodriguez	✓	✓		✓	✓	✓	✓			✓	✓	✓	✓	✓
Ernesto David Cancho-Rodriguez		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
Ricardo Richard		✓		✓	✓	✓	✓	✓		✓	✓	✓	✓	✓
Huamantango Navarro														

C : **C**onceptualization

M : **M**ethodology

So : **S**oftware

Va : **V**alidation

Fo : **F**ormal analysis

I : **I**nterpretation

R : **R**esources

D : **D**ata Curation

O : **O**riginal Draft

E : **E**diting

Vi : **V**isualization

Su : **S**upervision

P : **P**roject administration

Fu : **F**unding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data and the model developed in Python that support the findings of this research are available on GitHub at <https://github.com/martinmoralesb/ID9769>, reference number ID9769.




REFERENCES

- [1] K. Veena, K. Meena, Y. Teekaraman, R. Kuppusamy, and A. Radhakrishnan, "C SVM Classification and KNN Techniques for Cyber Crime Detection," *Wireless Communications and Mobile Computing*, no. 1, pp. 1–9, Jan. 2022, doi: 10.1155/2022/3640017.
- [2] A. Bilen and A. B. Özer, "Cyber-attack method and perpetrator prediction using machine learning algorithms," *PeerJ Computer Science*, vol. 7, pp. 1–21, Apr. 2021, doi: 10.7717/PEERJ-CS.475.
- [3] D. Wright and R. Kumar, "Assessing the socio-economic impacts of cybercrime," *Societal Impacts*, vol. 1, no. 1–2, pp. 1–4, Dec. 2023, doi: 10.1016/j.socimp.2023.100013.
- [4] G. Bhardwaj and R. K. Bawa, "Machine Learning Techniques Based Exploration of Various Types of Crimes in India," *Indian Journal of Computer Science and Engineering*, vol. 13, no. 4, pp. 1293–1307, Aug. 2022, doi: 10.21817/indjcs/2022/v13i4/221304142.
- [5] S. Morgan, "Cybercrime To Cost The World \$10.5 Trillion Annually By 2025," *Cybercrime Magazine*. [Online]. Available: <https://cybersecurityventures.com/hackerpocalypse-cybercrime-report-2016/>. (Date accessed: Jul. 16, 2024.)
- [6] N. AllahRakha, "Impacts of Cybercrimes on the Digital Economy," *Uzbek Journal of Law and Digital Policy*, vol. 2, no. 3, pp. 29–36, Aug. 2024, doi: 10.59022/ujldp.207.
- [7] A. G. Mohamed, A. Elsayed, and A. A. Galal, "Machine Learning for Detecting Cybercrime in the Banking Sector," *Journal of Southwest Jiaotong University*, vol. 58, no. 5, pp. 786–799, 2023, doi: 10.35741/issn.0258-2724.58.5.60.
- [8] A. Ampountolas, T. N. Nde, P. Date, and C. Constantinescu, "A machine learning approach for micro-credit scoring," *Risks*, vol. 9, no. 3, pp. 1–20, Mar. 2021, doi: 10.3390/risks9030050.
- [9] D. Taman, "Impacts of Financial Cybercrime on Institutions and Companies," *Arab Journal of Literature and Humanities*, vol. 8, no. 30, pp. 477–488, Feb. 2024, doi: 10.21608/ajahs.2024.341707.




- [10] C. Natarajan, S. D. Priya, and K. Isakkipriya, "Crime Rate Prediction Using Combined Arima and Lstm," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 6, no. 1, pp. 3694–3703, Feb. 2024, doi: 10.56726/irjmets48515.
- [11] N. S. Deepak, T. Hanitha, K. Tanniru, L. R. Kiran, N. R. Sai, and M. J. Kumar, "Analyze and Forecast the Cyber Attack Detection Process using Machine Learning Techniques," in *2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, IEEE, Jul. 2023, pp. 1732–1738, doi: 10.1109/ICESC57686.2023.10193289.
- [12] M. Vijayalakshmi and R. Norbu, "Smart Police: A Hybrid Deep Learning Model for Crime Proactivity Assessment," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, Jul. 2023, pp. 1–7, doi: 10.1109/ICCCNT56998.2023.10308231.
- [13] E. F. Aljarboua, M. B. Md. Din, and A. A. Bakar, "Cyber-Crime Detection: Experimental Techniques Comparison Analysis," in *2022 International Visualization, Informatics and Technology Conference (IVIT)*, IEEE, Nov. 2022, pp. 124–129, doi: 10.1109/IVIT55443.2022.10033332.
- [14] Y. Deng, H. Fan, and S. Wu, "A hybrid ARIMA-LSTM model optimized by BP in the forecast of outpatient visits," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 5, pp. 5517–5527, May. 2023, doi: 10.1007/s12652-020-02602-x.
- [15] D. Xu, Q. Zhang, Y. Ding, and D. Zhang, "Application of a hybrid ARIMA-LSTM model based on the SPEI for drought forecasting," *Environmental Science and Pollution Research*, vol. 29, no. 3, pp. 4128–4144, Jan. 2022, doi: 10.1007/s11356-021-15325-z.
- [16] S. Kulshreshtha and A. Vijayalakshmi, "An ARIMA-LSTM hybrid model for stock market prediction using live data," *Journal of Engineering Science and Technology Review*, vol. 13, no. 4, pp. 117–123, Aug. 2020, doi: 10.25103/jestr.134.11.
- [17] C. Kasemset, K. Phuruan, and T. Opasuwat, "Shallot Price Forecasting Models: Comparison among Various Techniques," *Production Engineering Archives*, vol. 29, no. 4, pp. 348–355, Dec. 2023, doi: 10.30657/pea.2023.29.40.
- [18] C. He, "A Hybrid Model Based on Multi-LSTM and ARIMA for Time Series Forecasting," in *2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP)*, IEEE, Apr. 2023, pp. 612–616, doi: 10.1109/ICSP58490.2023.10248909.
- [19] A. S. Temür and Ş. Yıldız, "Comparison of Forecasting Performance of ARIMA LSTM and HYBRID Models for The Sales Volume Budget of a Manufacturing Enterprise," *Istanbul Business Research*, vol. 50, no. 1, pp. 15–46, May. 2021, doi: 10.26650/ibr.2021.51.0117.
- [20] X. Liu, X. Xie, and Q. Guo, "Research on Cloud Computing load forecasting based on LSTM-ARIMA combined model," in *2022 Tenth International Conference on Advanced Cloud and Big Data (CBD)*, IEEE, Nov. 2022, pp. 19–23, doi: 10.1109/CBD58033.2022.00013.
- [21] A. S. Kuppli, "Forecasting Meteorological Variables and Anticipating Climatic Aberrations of an Oceanic Buoy Using A Neighbour Buoy," Doctoral dissertation, Dalhousie University, 2021.
- [22] S. Sharmin, F. I. Alam, A. Das, and R. Uddin, "An Investigation into Crime Forecast Using Auto ARIMA and Stacked LSTM," in *2022 International Conference on Innovations in Science, Engineering and Technology (ICISSET)*, IEEE, Feb. 2022, pp. 415–420, doi: 10.1109/ICISSET54810.2022.9775862.
- [23] A. F. Cunha, D. Ferreira, C. Neto, A. Abelha, and J. Machado, "A CRISP-DM Approach for Predicting Liver Failure Cases: An Indian Case Study," in *Lecture Notes in Networks and Systems*, vol. 271, pp. 156–164, 2021, doi: 10.1007/978-3-030-80624-8_20.
- [24] J. O. G. Jauregui, A. G. C. Cruzatti, M. A. C. Lengua, and H. V. Medrano, "Proposed Feature Selection Technique for Pattern Detection in Patients with Pneumonia Records," *International Journal of Online and Biomedical Engineering*, vol. 20, no. 7, pp. 69–89, May 2024, doi: 10.3991/ijoe.v20i07.47647.
- [25] L. Pahuja and A. Kamal, "EnLEFD-DM: Ensemble Learning based Ethereum Fraud Detection using CRISP-DM framework," *Expert Systems*, vol. 40, no. 9, pp. 1–18, Nov. 2023, doi: 10.1111/exsy.13379.
- [26] W. Safat, S. Asghar, and S. A. Gillani, "Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques," *IEEE Access*, vol. 9, pp. 70080–70094, 2021, doi: 10.1109/ACCESS.2021.3078117.
- [27] H. Wang and B. Zhang, "Research on ARIMA Model for Short-Term Traffic Flow Prediction based on Time Series," in *2023 8th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, IEEE, Nov. 2023, pp. 92–95, doi: 10.1109/ICIIBMS60103.2023.10347816.
- [28] J. Hernández, D. López, and N. Vera, "Primary user characterization for cognitive radio wireless networks using long short-term memory," *International Journal of Distributed Sensor Networks*, vol. 14, no. 11, pp. 1–20, Nov. 2018, doi: 10.1177/1550147718811828.
- [29] M. M. M. Barrenechea and M. A. C. Lengua, "A systematic literature review on the use of artificial intelligence for cybercrime rate forecasting," *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 3, pp. 2042–2054, Jun. 2025, doi: 10.11591/eei.v14i3.9213.
- [30] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, Jun. 2014, doi: 10.5194/gmd-7-1247-2014.
- [31] L. Borzemski and M. Wojtkiewicz, "Evaluation of chaotic internet traffic predictor using MAPE accuracy measure," in *Communications in Computer and Information Science*, vol. 160 CCIS, 2011, pp. 173–182, doi: 10.1007/978-3-642-21771-5_19.
- [32] Y. Zou, "The Prediction of Influenza Using the Hybrid ARIMA-LSTM Model," *Mathematical Modeling and Algorithm Application*, vol. 4, no. 2, pp. 20–25, Mar. 2025, doi: 10.54097/vgfxm178.

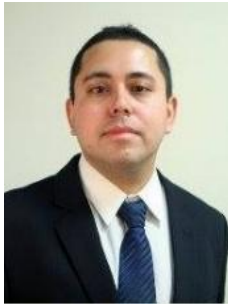
BIOGRAPHIES OF AUTHORS






Manuel Martin Morales-Barrenechea    is a senior professor at the Universidad Peruana de Ciencias Aplicadas (UPC). He holds a degree in Systems Engineering from UPC, a Master's in Business Administration and Management from Universidad Alas Peruanas (UAP) and he is currently candidate a Ph.D. in Systems and Computer Engineering at the Universidad Nacional Mayor de San Marcos (UNMSM). Professionally, he leads digital innovation initiatives in a prominent telecommunications company. He can be contacted at email: martin.moralesb@unmsm.edu.pe.






Ciro Rodriguez    is a senior professor and researcher at the Universidad Nacional Federico Villarreal and Universidad Nacional Mayor de San Marcos (UNMSM). Professor in Master's and Ph.D. postgraduate courses. He completed advanced training at the International Centre for Theoretical Physics (ICTP) in Trieste, Italy, and at the USPAS Particle Accelerator School in the United States. Additionally, he pursued studies focused on information technology policy development in South Korea. Senior member of IEEE. He can be contacted at email: crodriguezro@unmsm.edu.pe.



Ernesto David Cancho-Rodriguez    teaches at Universidad Nacional Mayor de San Marcos (UNMSM), School of Software Engineering. He holds a Global Master's in Business Administration (Global MBA). He holds a specialization in Business Data Analytics from The George Washington University (Washington, D.C., USA). His research interests include artificial intelligence, business intelligence, machine learning, and deep learning. He is currently a Ph.D. candidate in the Doctoral Program in Informatics and Systems Engineering. He can be contacted at email: ecanchor@unmsm.edu.pe.



Ricardo Richard Huamantingo Navarro    is a professional in Systems Engineering, holding a Master's degree in Systems Engineering with a specialization in Information Technology. He is currently pursuing doctoral studies as a Ph.D. candidate at the Universidad Nacional Mayor de San Marcos. He currently works as a programmer analyst in a public entity of Peru, designing, developing and implementing high-concurrency and scalability software at a national level. He can be contacted at email: Ricardo.huamantingo@unmsm.edu.pe.