INTERNATIONAL PAPER COMPETITION 2022
Department of Research and Technology
Ambarrukmo Tourism Institute (STIPRAM)
Yogyakarta, Indonesia
Jl. Ring Road Timur, Bantul, Yogyakarta 55584
Email: ristek@stipram.ac.id
Phone: +6281 229881414

INTERNATIONAL PAPER COMPETITION 2022

REPORTED MALICIOUS CODES INCIDENT WITHIN MALAYSIA'S LANDSCAPE: TIME SERIES MODELLING AND A TIMELINE ANALYSIS

Data Science

Muhammad Nadzmi Md Azam

Nor Azuana Ramli

Centre for Mathematical Sciences, College of Computing & Applied Sciences,
Universiti Malaysia Pahang, 26300 Gambang, Pahang, Malaysia.

nadzmimdazam@gmail.com

DEPARTMENT OF RESEARCH AND TECHNOLOGY
AMBARRUKMO TOURISM INSTITUTE (STIPRAM)
YOGYAKARTA, INDONESIA
2022

Reported Malicious Codes Incident within Malaysia's Landscape: Time Series Modelling and a Timeline Analysis

Muhammad Nadzmi Md Azam* and Nor Azuana Ramli

Centre for Mathematical Sciences, College of Computing & Applied Sciences,
Universiti Malaysia Pahang, 26300 Gambang, Pahang, Malaysia.

nadzmimdazam@gmail.com, azuana@ump.edu.my

*Corresponding Author
Received: day month 202x, Revised: day month 202x, Accepted: day month 202x
Published online: day month 202x

Abstract: The advancement of technology is such a marvel in these modern days. As countries embrace the vast progress of cyber-technology, the risk of cyber threats increases. Malicious codes have been one of the most menacing threats in the cyberspace. This research aims to investigate the outliers in the dataset timeline analysis. The data will be analysed to see the outliers and recognize what the crucial factor of the outliers in the data is. Then, the outliers will be investigated, and the findings will be constructed chronologically for the timeline analysis. The data also will be forecasted to predict the trend from May 2022 until December 2024. The predictive algorithms proposed are Autoregressive Integrated Moving Average (ARIMA), Long Short-Term Memory (LSTM), and NeuralProphet. The best model is chosen by the least values of mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). The outcome of this research is presented in an interactive dashboard as a deployment of this project. The results from the analysis showed that the best forecasting model is LSTM and from the forecasted data using this model, it can be seen the trend of incident increases until 2023, then decreases to 2024.

Keywords: Malicious Codes; Cybersecurity; Time Series; Forecasting; Long Short-Term Memory

1. Introduction

Technology and innovation have progressed to a level where men rely heavily on these things. These progressions have led us to the new wave of revolution, the Industrial Revolution. An industrial revolution is a term to illustrate technologies and innovation's role in improving and increasing the quality of life and the economy's sustained growth. It has been used since the 1800s, the year when the first industrial revolution started [1]. To this day, the term industrial revolution is used. Since 2021, the new wave of the industrial revolution is also known as Industrial Revolution 4.0 (IR 4.0). IR 4.0 is now affecting the growth of the economy through digitalisation.

Digitalisation in industries refers to using computers and automated systems to operate machinery or payment systems. IR 4.0, in simple words, is the revolution where industries use interconnectivity, automation, machine learning, and big data, whether offline or real-time [2]. A recent wave of revolutions means society needs to cater to the skills around IR 4.0. Hence, there will be rapid growth in cyberspace. Countries like Malaysia are starting to adopt IR 4.0 to sustain and improve their economy.

As the technological advances, the risk of cyber threats increases. Cyber threat is no longer a stranger thing in Malaysia. According to the Royal Malaysian Police, Malaysia has suffered a total cyber case of 11875, and the estimated loss is RM497,719 498. This number is predicted to increase in a worrying state of Malaysia. Cases like fraud, malicious codes, and intrusion are the major contributors to Malaysia's total cases [3]. Malicious code is the most damaging cyber threat. It can be defined as the codes or scripts that cause severe damage to a system, such as a computer. There are two classes of malicious codes which are independent and dependent. Independent, which can work on itself and dependent, which may need a trigger factor to activate. Malicious code, frequently known as malware, is the root of the most cyber threat in the world [4].

Thus, National Security Council has developed a strategy called Malaysia Cyber Security Strategy 2020-2024 (MCSS 2020-2024) as a guideline for handling and mitigating cybercrime in Malaysia. Building awareness and education about cybersecurity is one of the pillars of MCSS 2020-2024. Malware attacks especially the dependent one is triggered by human actions—for example, clicking on an unknown website. Hence, building a solid awareness of how malicious codes work is essential. This can be supported by timeline analysis and time series forecasting of the historical data reported malicious

code cases. Timeline analysis can help to investigate further outliers in the data distribution [5]. Meanwhile, time series forecasting can provide the trend of the studied subject [6], which is a reported incident for malicious codes. Timeline analysis and time series forecasting will be vital supporting facts in developing awareness for society as a plan to curb malicious code cases.

As an effort to align with the pillars of the strategy, research involving reported malicious codes in Malaysia's monthly data need to be conducted. Hence, this research was done with its main objective is to investigate the outliers in the dataset timeline analysis. To achieve the main objective, the dataset scrapped from the Malaysia Computer Emergency Response Team (MyCERT) official website were analysed to see the outliers and recognise the crucial factor of the outliers in the data. The data also were used to predict the trend from 2022 until 2024. The predictive algorithms proposed in this study are Autoregressive Integrated Moving Average (ARIMA), Long Short-Term Memory (LSTM), and NeuralProphet. The best model was chosen by the least values of mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). The outcome of this research is presented in an interactive dashboard as a deployment of this project. Thus, a good mitigation plan could take place by predicting the trend, building the timeline, and providing insight into the malicious code incident in Malaysia [6].

The remainder of the paper is organized as follows where in Section 2, some related studies were discussed to see the research gap between this study and previous studies. Then, Section 3 describes all the methodologies used to achieve all the objectives listed in this paper. Finally, the results obtained through analysis were discussed before the paper ended with a concise conclusion with some recommendations for future study.

2. Related Works

There are plenty of studies that had been done related to malware detection. Study done by [7] used data visualization in order to identify obfuscated malware. The novel detection method proposed in this study used hybrid models wherein static and dynamic malware analysis techniques were combined effectively along with visualization of similarity matrices to detect and classify zero-day malware efficiently. The results from this study showed that the proposed method provided high accuracy of classification with clear visualization since different malware families exhibit significantly different behaviour patterns.

Beside data visualization, data mining and machine learning techniques have provided promising results to detect hidden malware effectively in applications over a variety of platforms including smartphones and devices. Machine learning techniques such as support vector machine (SVM), Naïve Bayes, and k-nearest Neighbour (k-NN) have been applied in a number of studies to detect static malware. Study by [8] used different techniques in data mining and the results showed that SVM classifier achieved over 90% accuracy in detecting malware in mobile applications based on android platform.

Similar study with our research is done by [9] where in their study they proposed Long Short-Term Memory (LSTM) in detecting malware that focuses on run trace components in a dynamic analysis framework. The different with our study is their approach in detecting malware from the perspective of Natural Language Processing (NLP) by developing and testing models that process run traces of malicious and benign software. The results from the study showed that Instruction as a Sequence Model (ISM) achieved an accuracy of 87.51% and a false positive rate of 18.34%, while Block as a Sequence Model (BSM) achieved an accuracy of 99.26% and a false positive rate of 2.62%.

3. Material & Methodology

The Cross Industry Standard Process for Data Mining (CRISP-DM) methodology was utilised as the guideline for this research. The steps are research understanding, analytic approach, data requirement, data collection, data understanding, data preparation, modelling, evaluation, and deployment. The research flow of this methodology can be seen as in Figure 1 below:

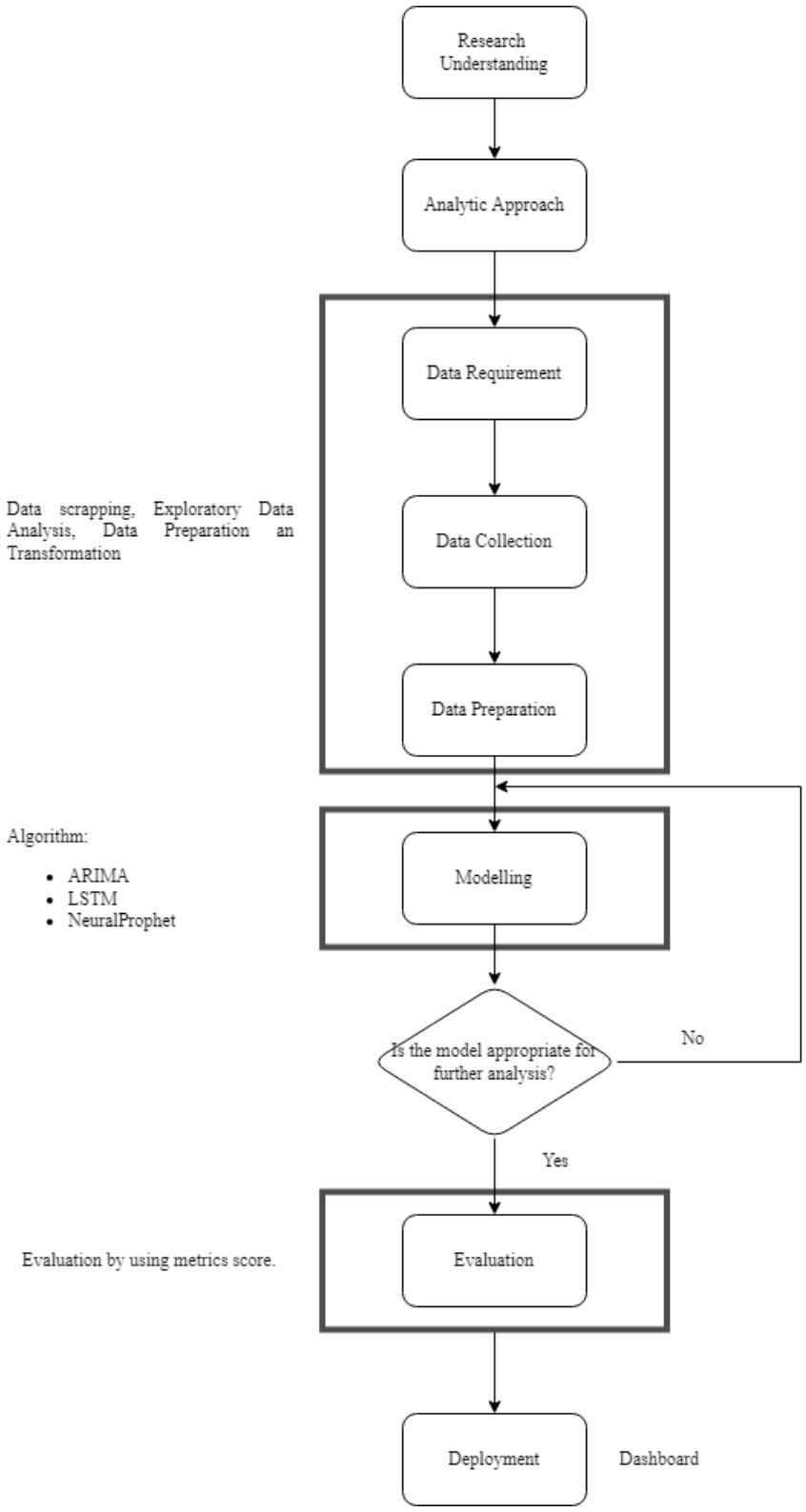

Figure 1. Research flow

3.1. Research Understanding

This research was conducted to answer some research questions such as:

i. Why does outliers' investigation important for a timeline analysis?
ii. What is the best time series forecasting model for this data?
iii. What is the type of deployment can be applied out of this project?

Questions mentioned the capability of log files in providing required data for this research. Throughout this research, time series modelling will be conducted, and a timeline analysis will be constructed.

3.2. Analytic Approach

For the analytic approach, this research adopted a predictive approach. The time modelling itself was developed to forecast data in the future.

3.3. Data Understanding and Data Collection

Data employed in the research was based on reported cases of malicious codes in Malaysia. The dataset is a monthly data, and it is univariate, where it only has one attribute with date as the index of the data. The dataset was scrapped from Malaysia Computer Emergency Response Team (MyCERT) official website which is https://www.mycert.org.my/portal/. From the website, the data was manually keyed in into a CSV file. Hence, there is no need to perform extract, load, and transform (ETL) process as it can be analysed on the go.

3.4. Data Preparation

Data preparation is a vital phase of this research because it helps to improve the models, to detect outliers, and to improve the quality the data. Firstly, exploratory data analysis was performed. This step is important in order to gain insights into our dataset. The data shape, missing values, and data type were checked. Then, all the results from the analysis were visualized in form of line plot, histogram, and box plot. Different plot was used for different analysis, as such the line plot was used to look for the trend and pattern of our data. Histogram was used to show the distribution of the data, and the box plot was used to recognise the outliers in the data. Once the outliers were recognised, then the analysis can be proceeded to the next phase.

Theoretically, outliers will affect the model quality. Due to that reason, outliers need to be treated by replaced with 25th quantile and 75th quantile depending on the values of the outliers. The distribution and box plot of the new dataset need to be checked before proceeded to the next step. The last step of data preparation is time series stationarity. When dealing with time series data, it is important for the data to achieve stationarity, where the mean, variance, and correlation are approximately constant over the time [10]. The stationarity of the data can be achieved by transforming and differencing depending on the condition of the data.

3.5. Modelling

Modelling phase will consider the best model for each algorithm. Parameter tuning is definitely involved in having the best models.

1) Autoregressive Integrated Moving Average (ARIMA)

Autoregressive integrated moving average (ARIMA) is a statistical analysis model that uses time series data to either better comprehend the dataset or forecast future trends. One can comprehend an ARIMA model by defining each of its components as AR, I, and MA. Autoregression (AR) refers to a model in which a changing variable regresses on its own lagged or previous values, while Integrated (I) denotes the differentiation of raw observations to permit the time series to become stationary and Moving Average (MA) integrates the relationship between an observation and a residual error from an MA model applied to lagging observations. Each ARIMA component serves as a parameter with a standardised nomenclature. A conventional notation for ARIMA models would be ARIMA with *p*, *d*, and *q*, where integer values replace the parameters to reflect the type of ARIMA model employed. *p*, *d*, and *q* may be used to define the parameters while *q* is the size of the MA window and sometimes known as the order of the MA. In this study, ARIMA will be the standard of forecasting algorithm. This is because of the nature of the algorithm that can predict data with a properly prepared data. ARIMA is said to be on par but conditionally compared to machine learning algorithm and deep learning algorithm depending on how the dataset were handled [11]. To find the best ARIMA model, stepwise search will be applied.

2) Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is a variety of recurrent neural networks (RNN). In a conventional RNN, there is a short-term memory. When paired with an LSTM, they have a long-term memory. The LSTM is an RNN extension that, in essence, extends memory. As a result, it is well adapted to learning from important events separated by considerable time gaps. Because of LSTM, RNN can remember inputs for a long time [12]. LSTM's ability to store information in a memory is comparable to a computer because LSTM can read, write, and delete information from its memory. LSTM is very reliable to forecast time series data. In LSTM, a cell, an input gate, an output gate, and a forget gate make up a typical LSTM unit. The three gates control the flow of information into and out of the cell, and the cell remembers values across arbitrary time intervals. Because there might be lags of undetermined duration between critical occurrences in a time series, LSTM networks are suitable for classification, processing, and making predictions based on time series data. The vanishing gradient problem that can occur when training traditional RNN was addressed by the development of LSTM.

3) NeuralProphet

Prophet is an open-source time series model generation algorithm created by Facebook that combines some classic principles with some novel twists. Working with time series data may be frustrating, and the many algorithms that build models can be picky and difficult to tune. This is especially true when dealing with data with different seasonality. Traditional time series models, such as SARIMAX, also contain several data constraints, such as stationarity and evenly spaced values. Other time series models, such as Recurring Neural Networks with Long-Short Term Memory (RNN-LSTM), can be quite complex and difficult to work with if the readers are unfamiliar with neural network architecture. As a result, time series analysis has a high entry hurdle for the average data analyst. Hence, in 2017, a group of Facebook researchers introduce the open-source project Facebook Prophet, which provides data analysts and data scientists with quick, powerful, and accessible time series modelling. It excels at modelling time series with numerous seasonality and avoids some of the limitations associated with other techniques [13]. Growth, seasonality, holidays, and error are the sum of three temporal functions plus an error term at its core. In 2021, Facebook again released an update to a version of prophet, which name NeuralProphet. NeuralProphet is said to bridge the gap between ARIMA and RNN model like LSTM. NeuralProphet is a strong algorithm when it comes to forecasting. The modular composability of the NeuralProphet model is a key idea. The model is made up of modules, each of which contributes to the forecast in some way. For a multiplicative impact, most components can be designed to be scaled by the trend. Each module has its own set of inputs and modelling methods. All modules, on the other hand, must provide *h* outputs, where *h* is the number of steps to be predicted into the future at the same time. The expected values, for the time series future values are added up. An arbitrary number of forecasts can be generated if the model is just time dependent. That exceptional situation will be regarded mathematically equal to a one-step ahead forecast with $h = 1$. The model components are:

$$\hat{y}_t = T(t) + S(t) + E(t) + F(t) + A(t) + L(t) \quad (1)$$

where $T(t)$ is trend at time, $S(t)$ is seasonal effects at time, $E(t)$ is event and holiday effects at time, $F(t)$ is Regression effects at time, $A(t)$ is Autoregression effect at time and $L(t)$ is Regression effect for lagged observation at time.

3.6. Evaluation

For model evaluation, this research will focus on evaluation using three metrics. The metrics are mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). The best model can be obtained based on the least values of these goodness-of-fit measured.

1) Mean Absolute Error (MAE)

The MAE is a statistic that measures the average magnitude of errors in a group of forecasts without taking their direction into account. It determines how precise continuous variables are. The equation is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y| \quad (2)$$

where *n* is the number of fitted point, $\hat{y}_i$ is the predicted value, and *y* is the observed value. In other words, over the verification sample, the MAE is the average of the absolute values of the differences between the forecast and the relevant observation. The MAE is a linear score, which means that all individual differences are equally weighted in the average.

2) Root Mean Squared Error (RMSE)

The RMSE is a quadratic scoring rule that calculates the average magnitude of an error. The RMSE equation is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (3)$$

To put it another way, the difference between the predicted, and the observed values is squared before being averaged throughout the sample. Finally, the square root of the average is determined. The RMSE gives enormous errors a lot of weight because the errors are squared before being averaged. As a result, the RMSE is most useful when errors are the least desirable. The MAE and the RMSE can be paired to diagnose forecast error variation. The RMSE will always be greater than or equal to the MAE; the larger the difference, the greater the variance in the individual errors in the sample. When the RMSE matches the MAE, all errors are of identical magnitude.

3) Mean Absolute Percentage Error (MAPE)

The mean absolute percentage error (MAPE), also known as the mean absolute percentage deviation, is used to assess the accuracy of a forecasting system. The formula is:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (4)$$

It is calculated as the average absolute percentage error minus actual values divided by real values for each time period, and it is expressed as a percentage. MAPE is the most used measure for forecasting error since the variable's units are scaled to percentage units, which makes it easier to understand. It's preferable if the data is free of outliers and presumably no zeros. It is frequently used as a loss function in regression analysis and model evaluation.

3.7. Deployment

The deployment where a dashboard was build based on the findings from the previous phase. The constructed timeline analysis will be included in the dashboard.

4. Results and Discussion

4.1. Result

All the results of this work are presented in this section.

4.2. Data Analysis

As explained in the previous section, the dataset was inspected and visualised in exploratory data analysis. Tableau and Python are both used as the tools to handle the dataset. Packages load for exploratory data analysis are pandas, numpy, and matplotlib. After the dataset was uploaded in Python, missing values and data type were checked. There are no missing values, and the type of data is correct. Then, the line plot for the data is plotted to observe the pattern of the data throughout the time. The plot is shown in Figure 2.

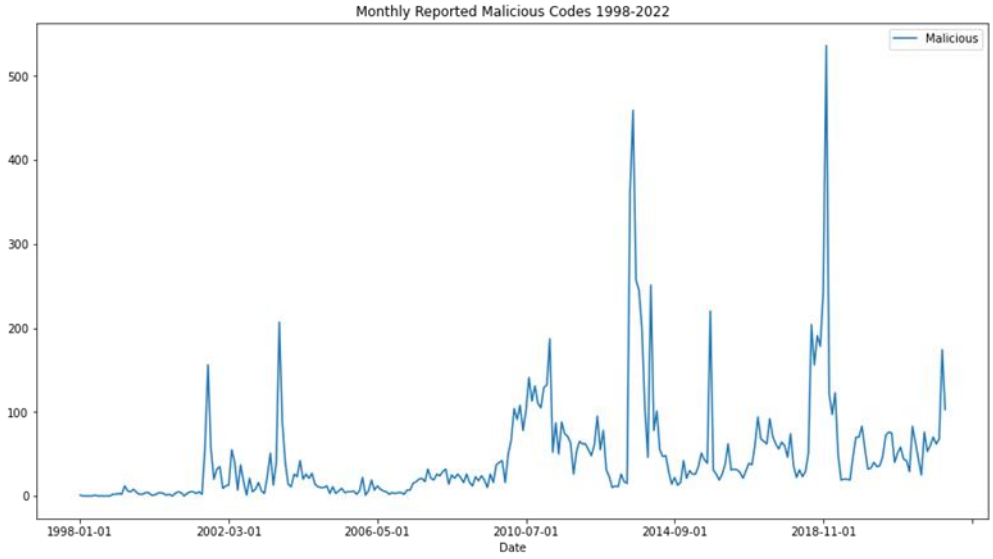

Figure 2. Line plot for the dataset.

Based on the visualised line plot, there might be an increasing trend in the line plot. However, there are huge spikes at certain data points, where it can be assumed that there might be outliers in the data distribution. From the visualisation, it can be hypothetically said that the dataset has a skewed data distribution and the outliers in the data set need to be investigated. To prove the assumption, a histogram and a box plot were plotted. Both plots are illustrated in Figure 3.

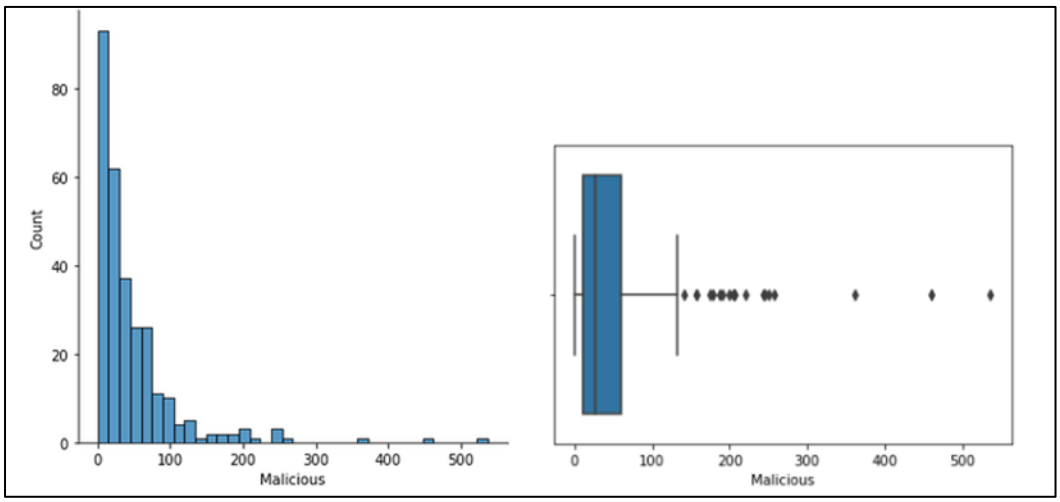

Figure 3. Histogram and box plot of the data.

The distribution of the data is positively skewed and, with the box plot shown, there are numerous outliers data point detected. The outliers will be treated by using the capping method. Capping method replaced outliers with the upper limit of the dataset where for this dataset, the upper limit data point is 136. After the outliers had been treated, the analysis is continued with the plotting of seasonal decompose plot. The plot is visualised as in Figure 4.

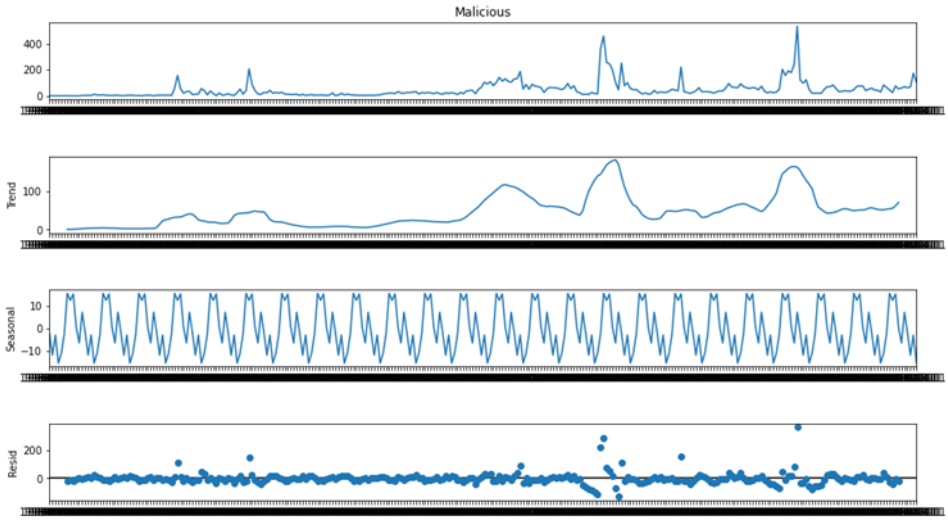

Figure 4. Seasonal decomposition.

Plot from Figure 4 showed that the trend is increasing, and there is seasonality in the dataset. Clearly, for certain years, there are highs and lows cases, thus, it proves that there is seasonality. However, the increasing trend of the dataset is not linear. It can be simplified that it can be seen the amplitude of the seasonal pattern periodically within a year. Hence, the analysis can be continued with transformation and differencing. These techniques will help with the stationarity of the data. First, a box cox transformation for variance, then the Augmented Dickey Fuller (ADF) Test will be run. The lambda, λ value obtained from the transformation is 0.2059, which is closer to 0. It means that it is ideal to do a log transformation on the dataset. After performing the log transformation, the λ value is 1.258 which is close to 0. The value indicated that the variance of the dataset is said to be stationary. Then, the analysis continued with performing the ADF test to provide the stationarity of the data in term of mean. The result of the test showed that p-value is equal to 0.04510 which is less than $\alpha = 0.05$. It can be concluded that the data is stationarity in mean. The result of the test unable us to skip the differencing phase because the mean is constant. Now the variance and the mean of the dataset are constant, the time series data is stationary, and the modelling shall commence.

4.3. Time Series Modelling

For the modelling, the data was partitioned into 80:20 set (80% training and 20% testing). Parameter tuning was conducted to obtain the best models, so the most optimised parameter for the models can be achieved. As it is confirmed that the data is stationary in term of mean and variance, ARIMA modelling can be performed. Important packages for the ARIMA were loaded to build the model. Stepwise searching was applied to find the best ARIMA(p,d,q) model. From the stepwise search results, the best model is ARIMA (2,1,3). The model has the lowest Akaike Information Criterion (AIC) which is 459.673. As the model is fitted with the best parameter, the model needs to be diagnosed and evaluated. To diagnose the model, the model residuals were plotted. The plots are visualised in Figure 5 below.

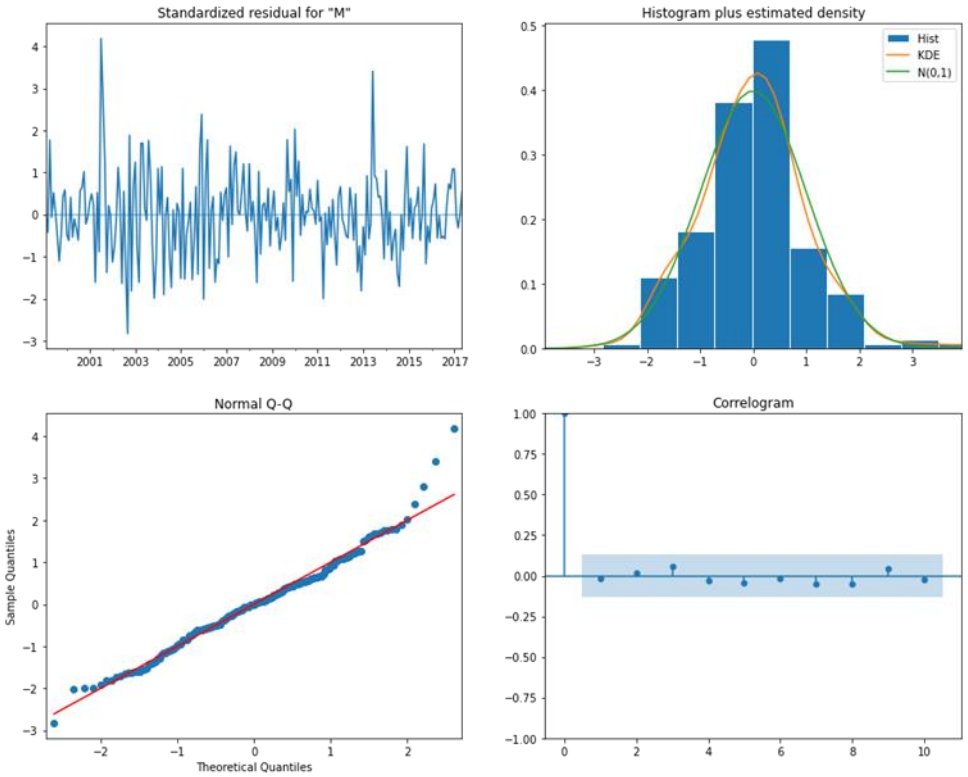

Figure 5. Model diagnostics.

From the normal Q-Q plot, it can be seen that the residuals are aligned on the theoretical, though some data points are astray. This suggests that no systematic departure from normality exists. The correlogram shows that there is no autocorrelation in the residuals, and they are effectively white noise. The model now needs to be evaluated using the three metrics and the value of the metrics are as shown in Table 1. Around 16.71% MAPE implies to the model, which means the model is 83.29% accurate to predict the next 59 data points. The MAPE indicates that ARIMA model is a suitable model for time series forecasting.

4.4. Long Short-Term Memory (LSTM)

For the LSTM modelling, Tensorflow and Keras library were used. The dataset was scaled using MinMax scaler before the model was built. This scaler was used to reduce the distance between the data point without changing the original data distribution. It also helps with the model building. In the model building phase, a stacked LSTM model was built with two layers of LSTM and two layers of dense. ADAM was applied for the optimizer and mean square error was applied for loss function. For the first model, the parameter setting was set to one epoch and one batch only. However, the model failed to provide best results as it did not have the slightest of the validation data pattern. To assure the model provide the best result, parameter tuning need to be done. The best epoch and batch for this model can be obtained by running models with different epoch and batch and measure it by using RMSE. Based on the parameter tuning, 500 epochs and one batch provided the lowest RMSE. The plot with the best parameter is shown as Figure 6.

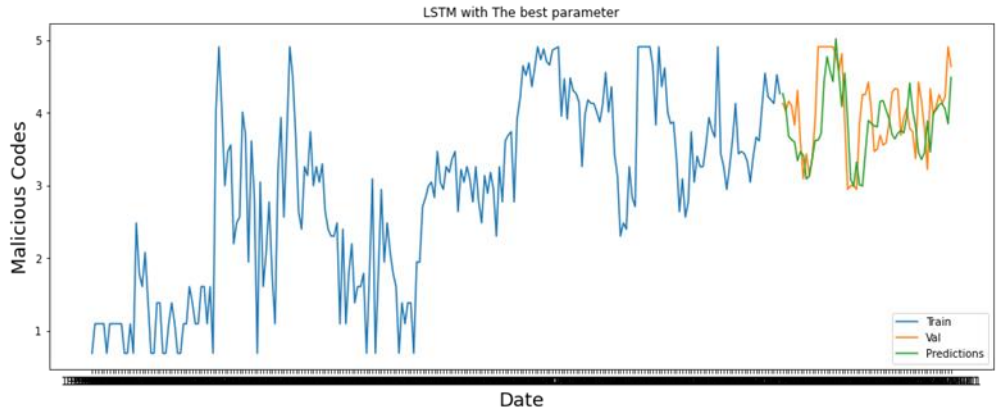

Figure 6. Model with parameter tuning.

Based on Figure 6, the prediction looks good as it has the validation data pattern and trend, even though it is not very identical. The evaluation of this model is tabulated in Table 1. Around 7.77% MAPE implies to the model, which means the model is 92.23% accurate to predict the next 59 data points. The MAPE indicates LSTM model is an excellent model for time series forecasting.

4.5. NeuralProphet

For the NeuralProphet, the batch was set to 16 and epoch was set to 306. The frequency set for the model is monthly, since our dataset is monthly. After running the model, the forecasted data and its decomposition were plotted. The plots are as shown in Figure 7.

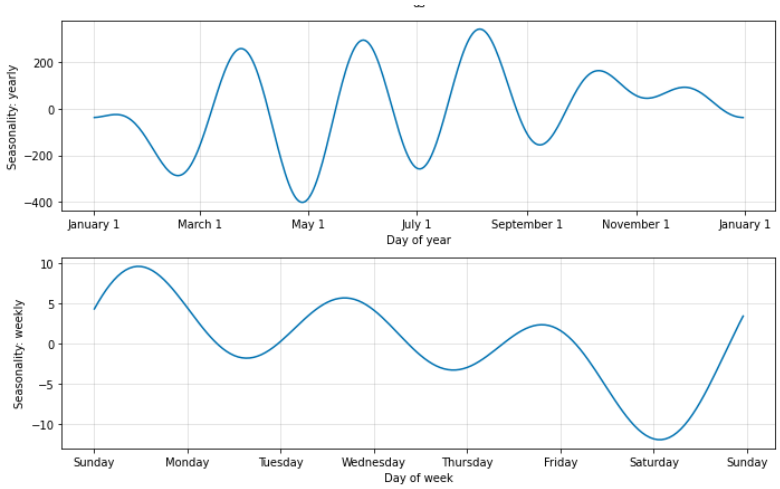

Figure 7. Forecasted model and its decomposition.

Based on the model, the forecasted data has a fluctuated pattern. Now the model is evaluated by using the proposed metrics. The result is tabulated as in Table 1. Around 24.71% MAPE implies to the model, which means the model is 75.29% accurate to predict the next 59 data points. The MAPE indicates that the NeuralProphet model is a suitable model for time series forecasting.

4.6. Comparison of the Models

Based on all the models proposed, it is confirmed LSTM performs the best, with ARIMA comes second. The evaluation of the model tabulated as in Table 1 below.

Table 1. Comparison of the models proposed.

Model	MAE	RMSE	MAPE
ARIMA	0.7922	0.9718	0.1671
LSTM	0.3038	0.4117	0.0777
Neural Prophet	0.9932	1.3218	0.2471

Besides comparison in terms of the metrics, the line chart with a prediction until December 2024 was also plotted for all models. The plot is visualized in Figure 8 below.

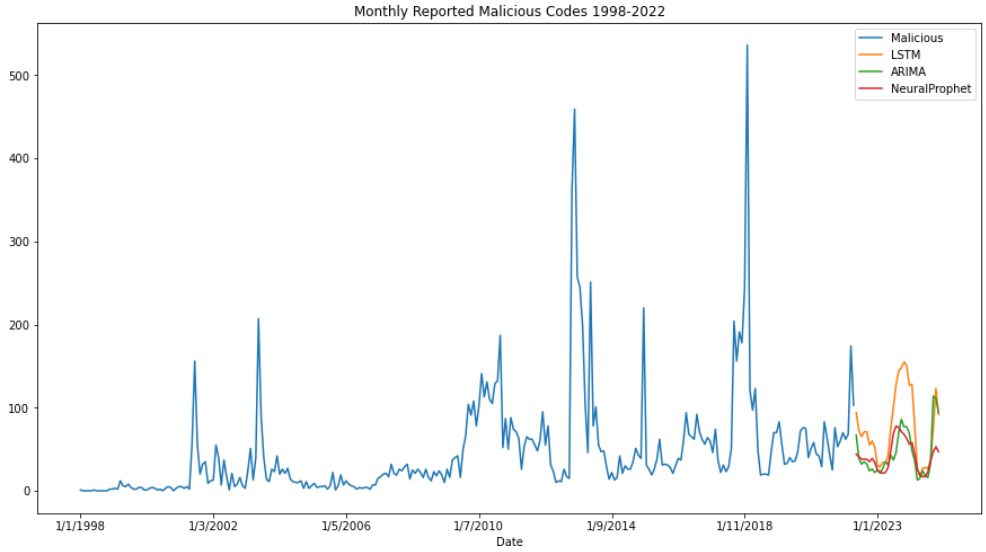

Figure 8. Line plot for an original dataset with forecasted data.

The model using LSTM shows a good trend and pattern. The highest cases prediction by LSTM will be in September 2023, and the lowest case prediction will be in May 2024. However, the forecasted data is not 100% reliable because of the stochastic nature of the dataset and unforeseen events that might occur in the future. Looking at the previous years, there are extreme highs in 2018, 2014, 2013, 2011, 2003, and 2001. These outliers can be explained by the constructed timeline analysis.

4.7. Timeline Analysis

For the timeline analysis, the outliers that were detected in data preparation phase were brought back to the dataset as it is important to further investigate these outliers. It will help the researcher and authorities to prepare and recognise current malicious codes threat. Timeline analysis also will help in developing a mitigation plan if there is a same event as these outliers happened in the future. The timeline analysis is illustrated as shown in Figure 9.

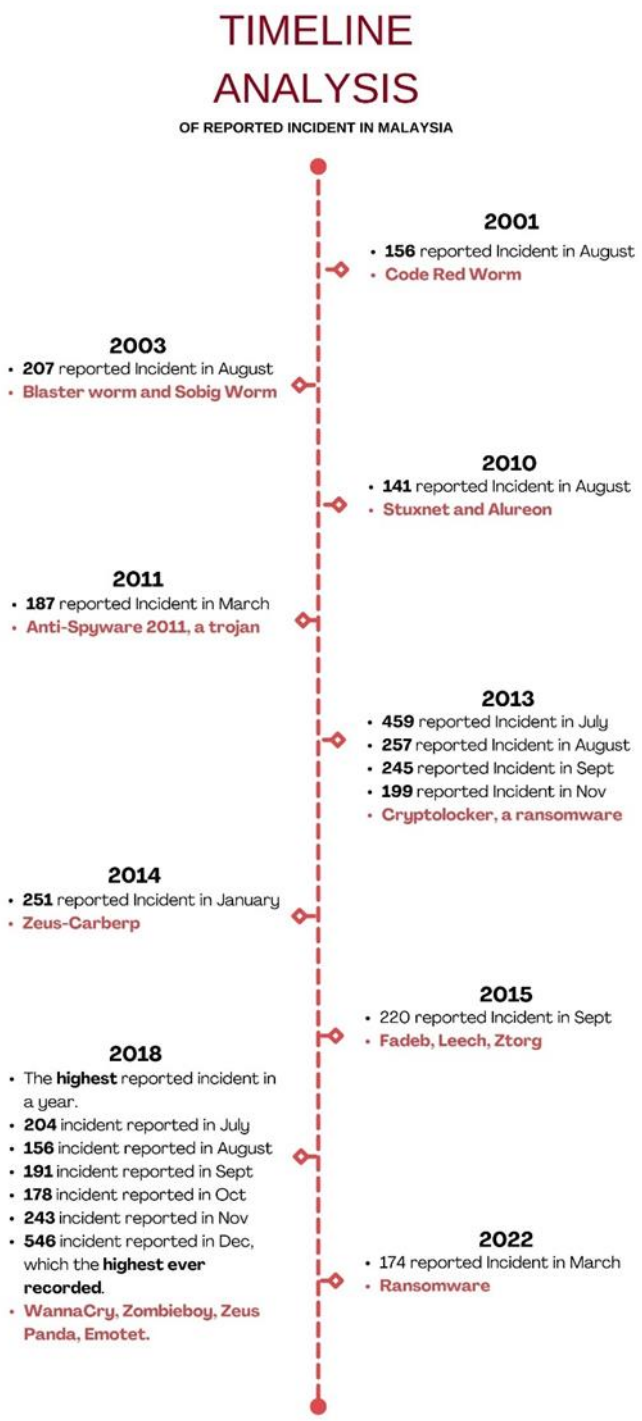

Figure 9. Timeline analysis infographic.

In 2001, the cases reported in August were 155 which is the most extreme data point in the year. The reason for these reported cases skyrocket in this year is because of the appearance of malicious codes, called Code Red Worm. Code Red and Code Red II worked by attacking through network and taking advantage of a flaw in Microsoft Web server Software [14]. Type of damage done by this worm is web defacement.

In 2003, the incidents reported in August were 207, which is the highest incident reported so far. The culprit was known as Blasterworm and Sobig Worm. These worms are lurking in the system similar to Code Red Worm. It exploited the security flaw within the Microsoft system. In 2010, the incidents reported in August were 141. The malicious codes behind the incident are Stuxnet and Alureon, where both of them are worms. The modus operandi is the same as other worms. These worms exploit zero-day vulnerability and can perform actions such as remote code execution.

In 2013, this year is the year of ransomware. Ransomware is a malicious code that encrypted your files in computers and demands a ransom or a payment to unblock it. Some ransomware threat will delete the files if the ransom is not paid. The most notorious ransomware in 2013 was Cryptolocker.

In 2014, the incident reported in January were 251. The main player of the cases is said to be Zeus-Carberp, a trojan that steals sensitive data. Zeus-Carberp is well known for cases such data breaches and data theft in banking institution. In 2014, the incident reported in September were 220. This is the year where mobile malware flourished. The most notable malwares are Fadeb, Leech, and Ztorg. These bad boys are ransomware.

2018 was the most miserable years for cyber space, yet a year where malicious codes incident thrived the most. The year has the most malicious reported incident cases a year ever recorded, where there were 548 incidents, the highest incident reported in month was recorded in December that years. A lot of key players contribute to the exponential increment of incident. The players are WannaCry, Zombieboy, Zeus, and Emotet. WannaCry contribute most of the reported incident [3].

Finally, a recent month in 2022 where 174 incidents were reported. The factor of the rising cases after so long is ransomware. To end the line, the timeline has been constructed and all the objectives for this research are achieved.

4.8. Deployment

For the deployment, a Tableau dashboard was created. The dashboard consists of line plot, line plot with forecasted data, heat map for total incident reported yearly, and timeline analysis. It will be published on the tableau public gallery and can be viewed at https://public.tableau.com/app/profile/nadzmi.azam/viz. The snapshot of the dashboard can be seen in Figure 10.

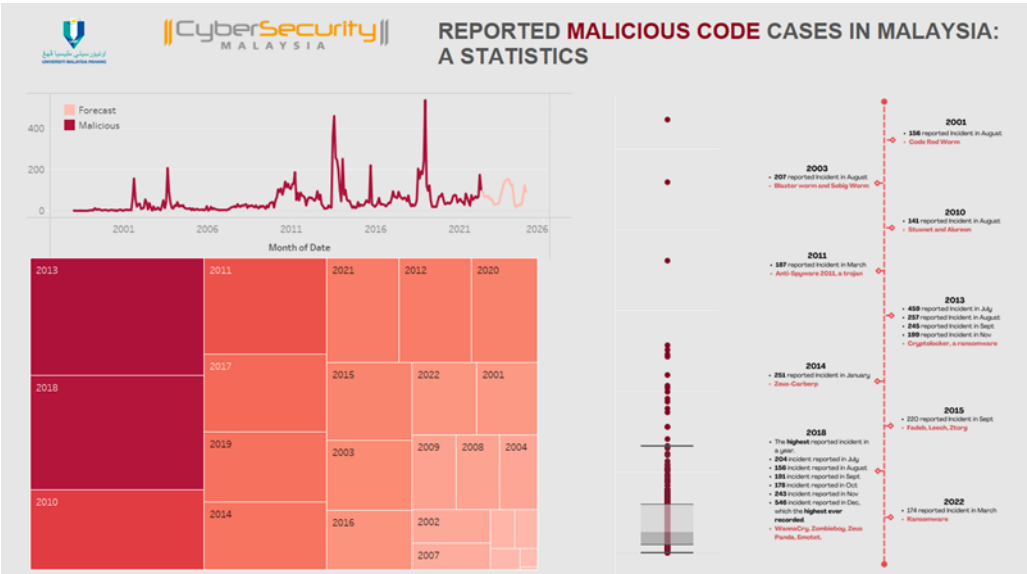

Figure 10. Deployment dashboard.

The dashboard is the most ideal deployment because it can be informative and useful to the mass on the reported malicious code incidents. Through this dashboard also we can solidify the awareness of cyber safety.

5. Conclusion

In summary, this project collected the univariate time series data for time series modelling and timeline analysis. The data was based on the reported malicious codes incident in Malaysia. During the data preparation, outliers data points were detected, which are needed for timeline analysis. Then, data was prepared to ensure the data is stationary, because stationarity in time series data is a necessity in dealing with time series data. After the data was prepared, the predictive modelling was done by using three different proposed algorithms which are ARIMA, LSTM, and NeuralProphet. The models were ensured to be the model respectively, so the best forecasting model can be obtained. All the models were evaluated using three proposed metrics which are MAE, RMSE, and MAPE. The best model obtained was LSTM, and the data was forecasted from May 2022 until December 2024. For the outliers, a lot of reading need to be done to recognise the factors of the outliers. From there, a timeline analysis with the outliers managed to be constructed. Finally, all the outcomes were presented in an interactive dashboard made with Tableau.

To conclude, throughout this research, all the objectives of the project were achieved. The questions of the research were also answered. The time series data can provide information for timeline analysis by searching the outliers in the dataset. The outliers are normally having a unique information because of the external and unforeseen event factors. Thus, this information can be used to construct a timeline analysis. The best forecasting model of this research is LSTM and from the forecasted data using this model, we can see the trend of incident increases until 2023, then decreases to 2024. With the summary and conclusion mentioned above, we can justify this research is successfully completed.

For the recommendations, we can tune the model more to improve the forecasting. However, these actions need more brains and muscles. Brains refer to individuals and muscles refer to the hardware. An outstanding model can be tuned with a greater understanding of the algorithm and dataset, and it is important to have a very good hardware such as good processing units but this costs a lot of capital. Other than that, the authority can use this proposed method to forecast other incident such as online frauds. Online frauds incidents have increased tremendously since the pandemic era. Thus, it is viable to use this proposed research to analyse and forecast the online fraud.

References

[1] Lucas Jr., R. E., "The Industrial Revolution: Past and Future," *Lectures on Economic Growth*, XLIV (8), 109–188 (2002).

[2] Khan, N., Khan, S., Tan, B. C., & Loon, C. H., "Driving Digital Competency Model towards IR 4.0 in Malaysia," *Journal of Physics: Conference Series*, 1793(1), 0–10 (2021). https://doi.org/10.1088/1742-6596/1793/1/012049

[3] MyCERT, Malaysia Threat Landscape 2018 - Based on Incidents Reported to MyCERT, https://www.mycert.org.my/portal/publicationdoc?id=270d8ee0-cdd1-49fb-827d-f8fca7752155 Retrieved 5 February, 2022.

[4] Splunk, Top 50 Security Threats, https://www.splunk.com/en_us/form/top-50-security-threats.html/ Retrieved 17 March, 2022.

[5] Studiawan, H., Sohel, F., & Payne, C., "Sentiment Analysis in a Forensic Timeline with Deep Learning," *IEEE Access*, 8, 60664–60675 (2020). https://doi.org/10.1109/ACCESS.2020.2983435

[6] Husák, M., Komárková, J., Bou-Harb, E., & Čeleda, P., "Survey of attack projection, prediction, and forecasting in cyber security," *IEEE Communications Surveys and Tutorials*, 21(1), 640–660 (2019). https://doi.org/10.1109/COMST.2018.2871866

[7] Venkatraman, S. and Alazab, M, "Use of Data Visualization for Zero-Day Malware Detection," *Security and Communication Networks*, Volume 2018, Article ID 1728303, 1-13 (2018).

[8] L.Sun,Z.Li,Q.Yan,W.Srisa-an,andY.Pan,"SigPID:significant permission identification for android malware detection," *In the Proceedings of the 2016 11th International Conference on Malicious and Unwanted Software (MALWARE)*, 1–8 (2016).

[9] C. Acarturk, M. Sirlanci, P. G. Balikcioglu, D. Demirci, N. Sahin, and O. A. Kucuk, "Malicious Code Detection: Run Trace Output Analysis by LSTM," *IEEE Access*, 9, 9625-9635 (2021). doi: 10.1109/ACCESS.2021.3049200

[10] Van Greunen, J., Heymans, A., Van Heerden, C., & Van Vuuren, G., "The prominence of stationarity in time series forecasting," *Journal for Studies in Economics and Econometrics*, 38(1), 1–16 (2014). https://doi.org/10.1080/10800379.2014.12097260

[11] Wang, F., Li, M., Mei, Y., & Li, W., "Time Series Data Mining: A Case Study with Big Data Analytics Approach," *IEEE Access*, 8, 14322–14328 (2020) https://doi.org/10.1109/ACCESS.2020.2966553

[12] Siami-Namini, S., Tavakoli, N., & Namin, A. S., "The Performance of LSTM and BiLSTM in Forecasting Time Series," *In the Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, 3285–3292 (2019). https://doi.org/10.1109/BigData47090.2019.9005997

[13] Taylor, S. J., & Letham, B., "Forecasting at Scale," *American Statistician*, 72(1), 37–45(2018). https://doi.org/10.1080/00031305.2017.1380080

[14] Technologist, C., "Testimony Before the Subcommittee on Government Efficiency, Financial Management, and Intergovernmental Relations, Committee on Government Reform, House of Representatives," *INFORMATION SECURITY Code Red, Code Red II, and SirCam Attacks Highlight Need for P*. (2001).