Forecasting the Air Pollution Index: A Case Study in Shah Alam, Selangor

Nur Hafiraniza Bakhtiar
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA,
Negeri Sembilan Branch, Seremban Campus, Negeri Sembilan, Malaysia
2020627664@isiswa.uitm.edu.my

Isnewati Ab Malek
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA,
Negeri Sembilan Branch, Seremban Campus, Negeri Sembilan, Malaysia
isnewati@uitm.edu.my

Haslinda Ab Malek
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA,
Negeri Sembilan Branch, Seremban Campus, Negeri Sembilan, Malaysia
haslinda8311@uitm.edu.my

Siti Sarah Januri
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA,
Negeri Sembilan Branch, Seremban Campus, Negeri Sembilan, Malaysia
sarahjanuri@uitm.edu.my

Jaida Najihah Jamidin
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA,
Negeri Sembilan Branch, Seremban Campus, Negeri Sembilan, Malaysia
jaida5698@uitm.edu.my

Article Info

ABSTRACT

The World Health Organization (WHO) defines air pollution as any chemical, physical, or biological agent that tampers with the atmosphere's natural characteristics and contaminates either the indoor or outdoor environment. The evaluation of air pollution can be done by using air pollution prediction. When air pollution levels are high, it can notify and warn the public while assisting the management of many different chemical compounds through policy. The objective of this study is to find the best forecasting model for the air pollution index (API). This study also attempts to predict the monthly mean concentration of the API in Shah Alam for 2023 by using the time series model. To achieve the objectives, the Box-Jenkins Methodology and Univariate Techniques were used. This study examines the API using Holt's Method, Double Exponential Smoothing Technique, and ARIMA models. Based on the smallest value of root mean squared error (RMSE) and mean absolute error (MAE), it shows that the most adequate model for the API for this period is the ARIMA model. Air quality forecasting is reliable and effective in controlling the composition of air pollution. With the ability to forecast the mean concentration of the Air Pollution Index, these findings could aid the Department of Environment in analyzing the substances that contribute to air pollution. Additionally, this information could help reduce the incidence of air pollution-related diseases among Malaysians.

Corresponding Author:
Isnewati Ab Malek
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Negeri Sembilan Branch, Seremban Campus, Negeri Sembilan, Malaysia
Email: isnewati@uitm.edu.my

PENERBIT PRESS
UNIVERSITI TEKNOLOGI MARA

1. Introduction

The Air Pollution Index (API) is employed in Malaysia to assess air quality. The API system indicated that air pollution consists of five main elements, which are ozone (O3), carbon dioxide (CO2), nitrogen dioxide (NO2), sulphur dioxide (SO2), and particulate matter (PM10). Malaysia was expected to become an industrialized nation in 2020, highlighting the urgent need to address the air quality issue that has been found to exacerbate existing respiratory health conditions [1]. Both natural and human activities contribute to air pollution in the country. Some of the natural causes of air pollution include soil dust, forests, and sea surface emissions. Commonly, these sources contribute to the level of air pollution. However, man-made sources such as the burning of fossil fuels, transportation emissions, and deforestation also contribute significantly to air pollution, apart from causing global climate change [2]. The air quality tended to be worsened due to the high level of man-made sources.

The Malaysian Department of the Environment (DOE) set up the Recommended Malaysian Air Quality Guidelines (RMG) in 1989 to control air pollutants. Subsequently, in 1993 the Malaysian Air Quality Index (MAQI) was introduced to inform the public about air quality conditions. In 1996, they revamped the system and adopted the air pollution index (API), which is modelled after the United States pollutant standard index (PSI) [3]. The API serves as an effective tool for assessing air quality. The API status indicator is categorized into five levels, including good, moderate, unhealthy, very unhealthy, and hazardous, as outlined in Table 1. These categories serve as benchmarks for air quality management and aid in interpreting data for decision-making processes.

Table 1. API Status

API Value	Status
0-50	Good
51-100	Moderate
101-200	Unhealthy
201-300	Very Unhealthy
>300	Hazardous

In this study, the API was utilized to determine the best model for forecasting the API from January 2023 to December 2023. Numerous studies showed major cities with high seasonal heating demands, heavy industry, and high vehicular traffic volumes, or with all these three, as the worst air pollution [4].

In 2022, 7.9 million people were living in Selangor. According to the Selangor State Structure Plan 2020, the state's population was anticipated to grow to 9 million people by the year 2035. Around 36,592.52 hectares of land had been designated for development, representing 80% of the total area. This indicates that the state struggled with air quality issues over the years due to population growth and development. Shah Alam is particularly susceptible to air pollution because of its densely populated surrounding areas, significant industrial and commercial developments, and heavy traffic [5]. These features make Shah Alam more vulnerable to air pollution [6]. Therefore, this study aims to identify the best forecasting model for the API and predict the API's monthly mean concentration in Shah Alam for 2023.

2. Literature Review

Air pollution is the presence of harmful substances in the atmosphere that exceed a certain concentration and cause negative effects on both, the ecological system and human life. With the growing concern for the environment, many researchers have conducted numerous studies, with air pollution forecasting being of utmost importance [7]. An accurate forecasting is a foundation for taking any effective pollution control measures, making it a crucial task.

Accurate AQI forecasting is crucial for safeguarding human health, protecting the environment, and supporting economic stability. It enables authorities to make decisions, facilitates sustainable urban planning, and aids in the management of pollution control strategies [8]. Recognizing the significance of the Air Pollution Index (API) and air quality forecasting, this study aims to predict the monthly mean API by using a time series model.

One of the most effective statistical techniques for forecasting from time-series data is the Autoregressive Integrated Moving Average (ARIMA) model, and often referred to as the Box-Jenkins model. ARIMA models are employed to identify the model that best fits the historical data in a time series. This forecasting model has been extensively used across various sectors without limitation to air pollutant time series [9]. Another type of time series model is univariate time series analysis, where a single variable varies over equal time increments in the data. The time increments can be daily, hourly, monthly, or yearly. Univariate modelling, known as a projective approach to forecasting, generates predicted values based on data from previous observations [10]. Like the Box-Jenkins methodology, univariate modelling has been widely used for forecasting in various industries, not just limited to air pollution time series.

In a study which focused on the Thiruvananthapuram District of Kerala, India, the ARIMA and Seasonal Auto-Regressive Integrated Moving Average (SARIMA) methods were employed to forecast air quality indices [11]. The study utilized monthly air quality data from 2012 to 2015 for nitrogen dioxide (NO2), sulphur dioxide (SO2), suspended particulate matter (SPM), and respirable particulate matter (RSPM), collected at four sites in Thiruvananthapuram District. The result indicated that all stations' air quality index readings between 2012 and 2015 fell within the satisfactory (51-100) AQI range. The researchers found that the ARIMA models outperformed SARIMA models in terms of forecasting accuracy.

Another study aimed to compare the performance of artificial neural networks (ANN) and Arima models for a better forecasting the air pollution data in Malaysia [12]. The outcomes highlight the fact that compared to ARIMA, the ANN provided the lowest forecasting error to predict API in Klang. As such, the ANN may be regarded as a reliable predictive method to generate data for the general public regarding the status of air quality at a particular time.

Additionally, a study utilized a time series technique with autoregressive integrated moving average (ARIMA) modelling to predict the maximum daily surface ozone (O3) concentration [13]. The study focused on surface O3 data that was collected at the airport in Brunei Darussalam between July 1998 and March 1999. The fitted ARIMA model had an order of (1,0,1), and it was found that the maximum O3 concentrations predicted by the model closely matched the observed values. The model's effectiveness was evaluated with several widely used statistical metrics.

In a study conducted by [14], it was about the air quality index (AQI) in Miyun County, Beijing, China. The original AQI data were found to be non-stationary during the model construction process. However, the first-order differencing data of the original AQI data were stationary. After comparing various models, the ARIMA (3,1,3) model was selected as the final model for fitting the ARIMA model. The least squares approach was used to represent the data in the Holt exponential smoothing model fitting. In terms of capturing trends and minimizing mean squared error (MSE), Holt modelling outperformed ARIMA modelling in these two model fittings. Therefore, based on this data, the Holt model is preferred for predicting future AQI values.

Air pollution levels reflect the environment's health, and declining air quality immediately affects public health. Air quality forecasting, monitoring, and early warning systems are essential preventive measures for sustainable smart cities, environmental sustainability, and pollution control management [2]. Forecasting air quality is effective in protecting public health by giving early warnings about dangerous air pollutants [7]. Moreover, it aids the Department of Environment (DOE) in planning future monitoring programs and identifying significant air quality changes that could harm the environment and human health. API forecasting enables real-time processing and analysis of air quality data at the network's edge, providing decision-makers with timely and accurate information to address air pollution effectively [15].

Hence, it is clear that modelling and forecasting the Air Pollution Index (API) can be beneficial for various organizations. Therefore, the objective of this research is to develop a model using API time series data from Shah Alam, Selangor. This model aims to provide estimation of future API values, which can be valuable for understanding and managing air quality in the city.

3. Methodology

3.1 Description of Data
This study used secondary data that was extracted from Malaysia's Department of Environment (DOE). The monthly mean concentration of Air Pollution Index (API) of all five main elements which were the particulate matter with a diameter of 10 micrometres or less (PM_{10}), ozone (O_3), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), and carbon dioxide (CO_2) from January 2012

until December 2022 was obtained from Malaysia's Department of Environment (DOE) website. A total of 132 months of data on the API value in Shah Alam from January 2012 to December 2022 was applied to forecast the future of Shah Alam's API value.

3.2 Box-Jenkins Methodology

To achieve the objectives of this study, Box-Jenkins were used. The Box-Jenkins method comprises three primary models: autoregressive (AR), integrated (I), and moving average (MA). The AR and MA models are appropriate for analyzing stationary time series patterns. The combination of AR and MA models results in ARMA models. In cases where the data is non-stationary, the *I* model is used to transform the dataset into a stationary form, allowing for the creation of ARIMA models [16].

i. Autoregressive Integrated Moving Average (ARIMA) Model

To better comprehend or predict future values of time series, one could utilize an autoregressive integrated moving average model, or ARIMA model. The Box-Jenkins (1970) ARIMA model was a type of regression analysis that was capable in evaluating the strength of the dependent variable to its independent variable. It had three major processes which were an autoregressive (AR) of order *p*, differencing of degree *d* to render the time-series stationary, and moving average (MA) of order *q*. It was abbreviated as ARIMA (*p*,*d*,*q*) [17]. A simple model case ARIMA (1,1,1) was as shown below,

$$w_t = \mu + \emptyset_1 w_{t-1} - \theta_1 \varepsilon_{t-1} + \varepsilon_t \tag{1}$$

where $w_t = y_t - y$ serves the first difference of the series and was considered to be stationary. In this scenario, the values of *p*=1, *d*=1 and *q*=1. The equation could also be written as,

$$(1 - \emptyset_1 B)w_t = \mu + (1 - \theta_1 B)\varepsilon_t \tag{2}$$

The ARIMA model was applied by using the Box-Jenkins framework, which assumes stationarity in the data series. Stationarity means that the statistical properties of the data, such as the mean and variance, remain constant over time. This assumption implies that all instances of the process, regardless of when they occur, exhibit the same statistical characteristics [18].

Differentiation is a method used to transform non-stationary time series data into stationary form. This process involves some calculations of the differences between consecutive observations, which helps stabilize the mean of the series by removing trends and seasonality. Specifically, the first difference is computed as $W_t = y_t - y_{t-1}$, where W_t represents the difference between the current API value, y_t and the previous API value, y_{t-1}. If the first difference indicates that the data series is non-stationary, thus, this transformation is applied to achieve stationarity.

ii. Stages in ARIMA Model Development

The Box-Jenkins modelling approach had four main stages: Model Identification, Model Estimation and Validation, and Model Application as presented in Figure 1.
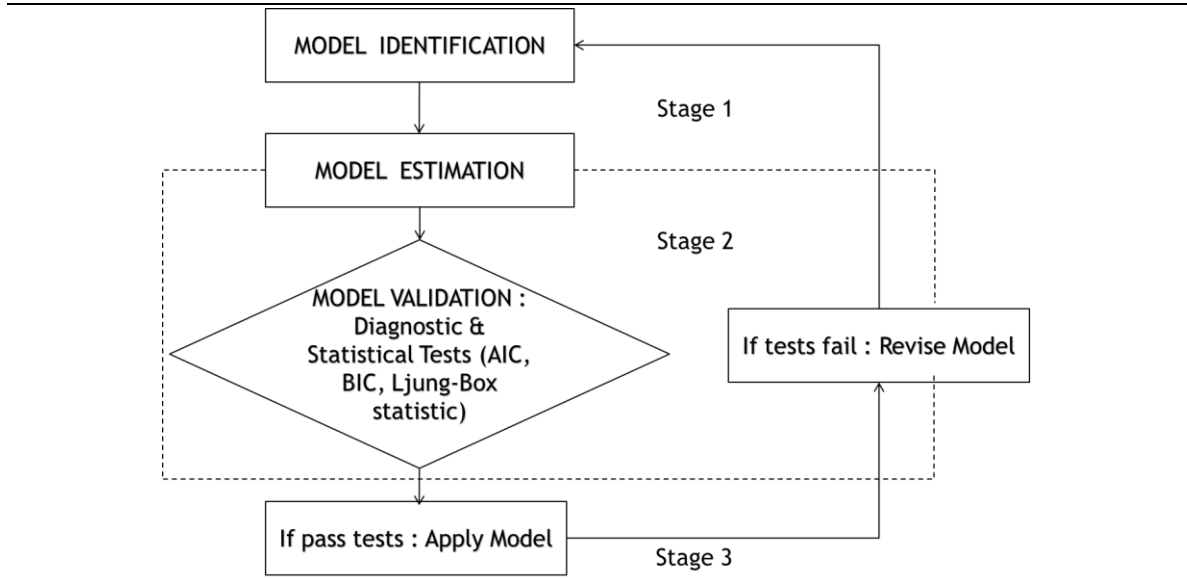
Figure 1. The Stages in ARIMA Model Development

In the Box-Jenkins technique, model identification is the initial step, which involves determining the most suitable class of models for the dataset. Once the data is made stationary, the parameters of the ARIMA model need to be determined. The ARIMA model employed three coefficients, p, d, and q, where p was the number of autoregressive terms, q denoted the number of moving average terms, and d specifies the order of differencing needed for stability. To find the autoregressive (p) and moving average (q) parameters, the autocorrelation function (ACF) and partial autocorrelation function (PACF) are used. The number of spikes in the PACF indicates p, whereas the number of spikes in the ACF determines q [19].

The next step is model estimation and validation. The ARIMA model was implemented using R-software, which was also employed for forecasting the monthly mean concentration of the Air Pollution Index (API) in Shah Alam, Malaysia. To validate the ARIMA models, statistical measurements such as the Ljung-Box Statistic, Akaike's Information Criteria (AIC), and Bayesian Information Criterion (BIC) were utilized.

The serial correlation of the residuals was assessed by using the Ljung-Box statistic, which helps in determining the adequacy of the model and the randomness of the residuals [20]. The hypothesis of the test is as follows:

H_0 : The errors are random (errors are white noise)

H_1 : The errors are non-random (errors are not white noise)

If the probability value is less than 0.05, the null hypothesis would be rejected, indicating that the model is mis-specified or inadequate. Conversely, if the probability value is more than 0.05, suggesting that the model is adequate.

The AIC and BIC are commonly used to evaluate the fitness of ARIMA. AIC is utilized to analyze various potential models and determine the one that best fits the data. On the other hand, BIC aims to balance model complexity and goodness of fit to produce the most accurate out-of-sample forecast. When both values are low, both criteria show that a model is the best ARIMA model. One similarity between the AIC and BIC was that the model was considered as the best ARIMA model when both values were low [21]. The formula of AIC and BIC are as follows:

$$AIC = e^{\frac{2k}{T}} \frac{\sum_{t=1}^{T} e_t^2}{T} \tag{3}$$

$$BIC = T^{\frac{k}{T}} \frac{\sum_{t=1}^{T} e_t^2}{T} \tag{4}$$

where T is the number of observations and k is the estimated model's total number of parameters, including the constant.

Once all test conditions are met and the model's fitness is confirmed, it can be applied to obtain forecast values. These values can be represented using confidence intervals or single-value estimates. Confidence interval estimation provides a valuable stochastic measure of the certainty and uncertainty associated with the forecasted values. However, if the test conditions are not satisfied, the model needs to be revised.

3.3 Univariate Techniques

The term univariate solely refers to the forecasts that relies on a sample of time series data for the air pollution index, without any consideration about the influences of other variables such as air temperature and wind direction. In this section, the application of Holt's method and Double Exponential Smoothing method was explained in this study. Consequently, the notations which are utilized in this research are denoted by y_i and t, representing the API and month, respectively.

i. Holt's Method

This technique is commonly used to handle data with linear trends because it offers more flexibility in tracking trends and slopes at different rates, besides smoothing them directly using multiple constants. In Holt's method, a final forecast was created by combining three primary equations: an exponential smoothing equation, a trend smoothing equation, and a forecast equation.

The exponentially smoothed series,
$$S_t = \alpha(y_t) + (1 - \alpha)(S_{t-1} + T_{t-1}) \tag{5}$$

The trend estimate,
$$T_t = \beta(S_t - S_{t-1}) + (1 - \beta)T_{t-1} \tag{6}$$

Forecasts m period into the future,
$$F_{t+m} = S_t + T_t \times m \tag{7}$$

where, S_t = smoothed value, S_{t-1} = the smoothed value for the previous period, y_t = smoothed constant for trend estimate ranges from 0 to 1, T_t = trend estimate, T_{t-1} = the trend estimated for the previous period, m = period to be forecast into future, which is 12 months, F_{t+m} = the forecast for the m periods into the future.

ii. Double Exponential Smoothing

Double exponential smoothing is also known as Brown's method. It is useful for a series that has many traits of a linear trend. The primary benefit of double exponential smoothing utilization over single exponential smoothing is the ability to produce multiple forecasts for the future. Generally, four main equations were involved. The notations are used to illustrate this technique as follows:

Let S_t represent the exponentially smoothed value of y_t at time t, and S_t' denote the double exponentially smoothed value of y_t at time t.

The single exponentially smoothed of API,
$$S_t = \alpha(y_t) + (1 - \alpha)S_{t-1} \tag{8}$$

The double exponentially smoothed value of API,
$$S_t' = \alpha(S_t) + (1 - \alpha)S_{t-1}' \tag{9}$$

The difference between both of exponentially smoothed values,
$$\alpha_t = 2S_t - S_t' \tag{10}$$

The adjustment factor which was forecast for one-year steps ahead,
$$F_{t+m} = a_t + b_t m \tag{11}$$

where F_{t+m} forecast the API at period m made in period t, for $m = 1,2,3, \dots ,12$ and $t = 1,2,3, \dots ,132$.

The accuracy prediction of the model, which was provided by most statistical software, has been compared using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

6

$$RMSE = \sqrt{\frac{\sum_t^n e_t^2}{n}}$$
$$MAE = \frac{1}{n}\sum_{t=1}^{n}|(y_t - \hat{y}_t)|$$

which $e_t = y_t - \hat{y}$, where y_t is the actual value at time t and $\hat{y}$, was the fitted value at time t.

4. Results and Discussion

4.1 Trend of Air Pollution Index (API)

As illustrated in Figure 2, a drawing of time series plot is the initial stage in time series analysis and gives a rough knowledge of the time behaviour of the series. The original series trend seems to have slightly decreased. However, this needs to be verified and proven by using descriptive analysis and trend modelling.
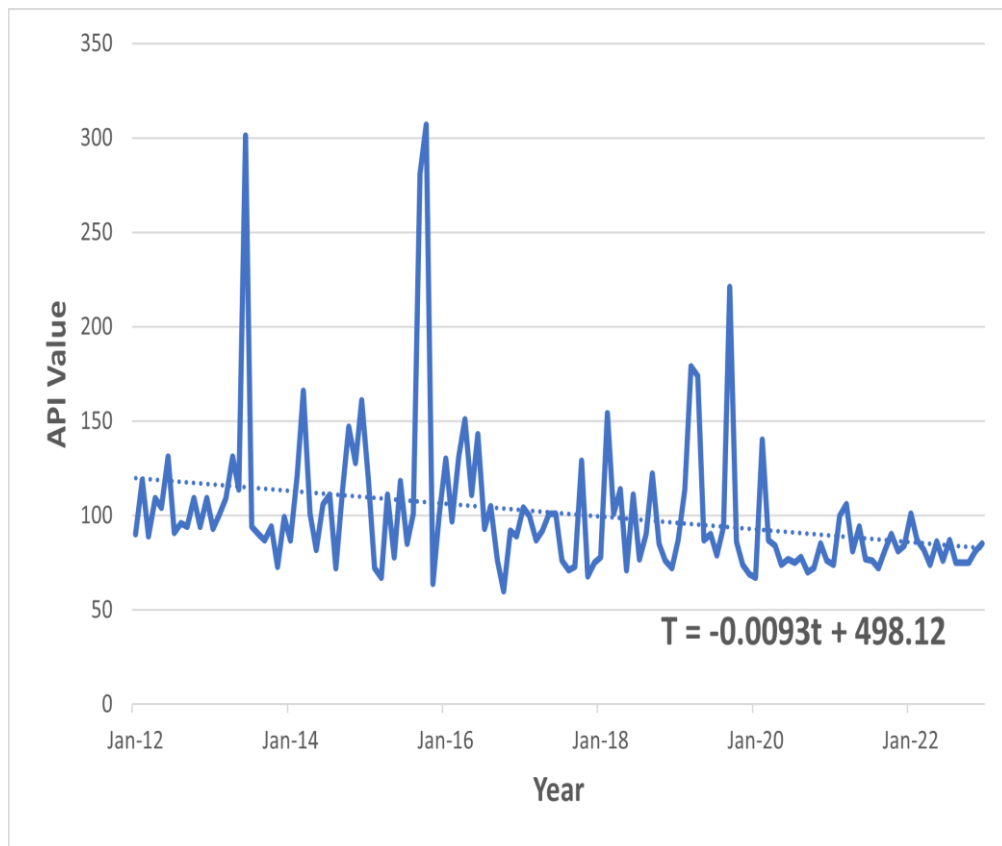
Figure 2. Air Pollution Index (API) Trend Line Graph

Figure 2 illustrates the monthly time series of the Air Pollution Index (API) in Shah Alam, Malaysia, from January 2012 to December 2022. The monthly API fluctuates throughout 132 months with the trend line, T = -0.0093t + 498.12. This equation indicates that as time increases, the API trend decreases by 0.0093. Analysis indicates that the air monitoring station at TTDI Shah Alam in Selangor recorded the highest API reading of 302 in June 2013. API levels exceeding 300 are considered hazardous. In that month also, the monitoring station at Muar experienced its worst, with 663 (emergency level) air quality readings due to haze episodes. Additionally, other locations recorded unhealthy, very unhealthy, and hazardous API levels, indicating widespread poor air quality [22].

Air pollution is expected to be worse during the dry season (June-September) than the wet season (November-March) because there is less rain. More rain during the wet season is expected to improve air quality, as it washes out and reduces the concentration of air pollutants in the

7

atmosphere [23]. However, in 2015 the highest peak in API records was observed in October, with a hazardous reading of 307. The trend of API good days showed a peak during November–December, which was during the wet season (Figure 2). Meanwhile, the average trend of API unhealthy days also showed the peak during the wet season, which was in February. A downtrend component is evident from the end of 2019 to the end of 2022, likely due to the temporary closure of factories in Shah Alam during the COVID-19 pandemic. Additionally, a turning point occurs from the middle of 2015 to the beginning of 2016, where the API value transitions from a downward trend to an upward trend.

4.2 Box-Jenkins Methodology

In the Box-Jenkins methodology, it is assumed that the characteristics of the initial data series are known. The fundamental assumption is that the data series is stationary. If the series is not stationary, differentiation is necessary to achieve stationarity before proceeding to the next stages.

Figure 3 revealed a stationary and a constant mean in the series. The fluctuation of the line around zero demonstrates its stationary nature. Since the probability value was less than 0.05, the Augmented Dickey-Fuller Test (ADF) revealed that the series was stationary (ADF test statistic = -4.3392, p-value = 0.01). Thus, the series does not require in differencing, and it is said to be integrated of order zero $[I(0)]$. The model obtained is represented in general term as ARIMA (p,d,q) with d=0.

ACF for Air Pollution Index
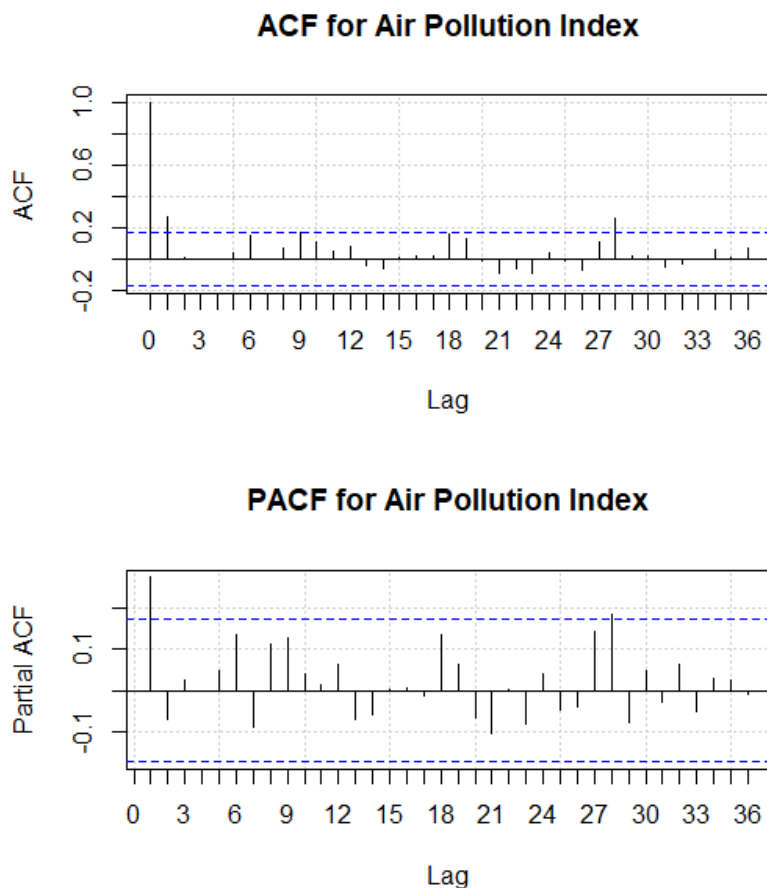

PACF for Air Pollution Index

Figure 3. ACF and PACF of Air Pollution Index (API)

Significant spikes were observed respectively in the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) for the parameters q and p. The confidence limits in this study were determined to be (-0.174078, +0.174078), although these limits can be varied based on the parameters of the ARIMA models. Identifying the right model formulation was rather difficult

because of the nature of the economic or business data series. Hence, several models to be the best possible formulations were identified and estimated.

In Figure 3 (ACF), two significant spikes were observed at lag 1 and 9 that determine MA(q=2). Meanwhile, Figure 5 (PACF), showed only one significant spike at lag 1, indicating AR(p=1). However, these observations do not guarantee the correct values of p and q for the model. Hence, several model formulations would be estimated in ensuring that a well-specified model was not missed out. The five specified models were ARIMA (2,0,2), ARIMA (1,0,2), ARIMA (1,0,1), ARIMA (1,0,3) and ARIMA (0,1,2).

AIC and BIC were used to assess and select the best ARIMA models, with the lowest values indicating the preferred models. To determine whether a serial autocorrelation occurs in a time series or vice versa, the value of the p-value for each model for the Ljung-Box test statistic was examined. The null hypothesis of the Ljung-Box Test assumes that the results exhibit white noise properties, which indicate no significant serial autocorrelation and satisfy the stationary condition. A model was considered to have no serial autocorrelation if the p-value of the Ljung-Box Test was greater than 0.05. The best model was selected by considering the absence of serial autocorrelation along with other relevant factors. Table 2 displays the AIC, BIC, and p-value of the Ljung-Box test for the selected model.

Table 2. Summary of the Estimated Model

Models	AIC	BIC	Box-Ljung (p-value)
ARIMA (2,0,2)	1032.819	1048.390	0.9980
ARIMA (1,0,2)	1030.926	1043.902	0.9978
ARIMA (1,0,1)	1029.539	1039.920	0.9558
ARIMA (1,0,3)	1032.858	1048.429	0.9983
ARIMA (0,1,2)	1331.503	1340.128	0.9370

Based on the results in Table 2, the p-values of the Ljung-Box test for the selected models were all above 0.05. This indicates that the errors for each model can be adequately represented by white noise, leading to the acceptance of the null hypothesis and suggesting no serial correlation in the model. The best model, ARIMA (1,0,1), was determined from Table 2 as it exhibited respectively the lowest AIC and BIC values of 1029.539 and 1039.920. Additionally, as the Principle of Parsimony favors choosing the simplest model, thus, ARIMA (1,0,1) was deemed the most appropriate selection. Therefore, ARIMA (1,0,1) is recommended as the most accurate model for predicting the Air Pollution Index (API) in Shah Alam for 2023.

4.3 Univariate Techniques

The API data was analyzed using univariate techniques, specifically Holt's method and the double exponential smoothing technique. The goal was to determine the best model, and the results were compared, as shown in Table 3.

Table 3. Error Measure of Holt's Method and Double Exponential Smoothing

Model / Error Measures	RMSE	MAE
Holt's Method	9.9964	8.2513
Double Exponential Smoothing	10.4605	8.2588

The optimal parameters for both models were determined using Excel software. In Holt's method, the parameter values for alpha and beta were respectively found to be 0.5237 and 0.01. Meanwhile, the optimal parameter for alpha was determined to be 0.2473 in the double exponential smoothing technique.

Based on the error measures using these parameter values, Holt's method was deemed more effective for forecasting the API compared to the double exponential smoothing technique. This conclusion was reached by comparing the evaluation components of RMSE and MAE, as shown in Table 3. Therefore, Holt's method is recommended as the best model among the univariate techniques for forecasting the API.

4.4 Determining the Best Model

The reliability of a model is not determined by the existence of a single good forecast. A good forecast model consistently produces good forecast values. Forecasters commonly assess a series of errors that have been produced over time or across time series before making a judgement about the model's quality. A model is therefore deemed superior to others if it meets a set of criteria. It is more customary to measure the error with the smallest magnitude. To determine the most effective method for forecasting the API, the performance of the Box-Jenkins model and univariate techniques were compared, as detailed in Table 4.

Table 4. Best Model Selection

Model	RMSE	MAE
ARIMA (1,0,1)	8.3481	6.7439
Holt's Method	9.9964	8.2513

The evaluated final models were ARIMA (1,0,1) and Holt's method. The error measures for ARIMA (1,0,1) were lower than those for Holt's method, as indicated in Table 4. Specifically, an analysis between the RMSE and MAE values from the evaluation section showed that ARIMA (1,0,1) had much lower values, with 8.3481 and 6.7439, respectively. Therefore, ARIMA (1,0,1) was selected as the best model for forecasting the API in Shah Alam.

4.5 Forecast the Air Pollution Index (API) in Shah Alam

Since ARIMA (1,0,1) model was chosen as the best model for forecasting the API in Shah Alam, it was used to forecast the data for the next 12 months. The future trend forecast is depicted in Figure 4, which appears to be satisfactory at 95 percent confidence interval.
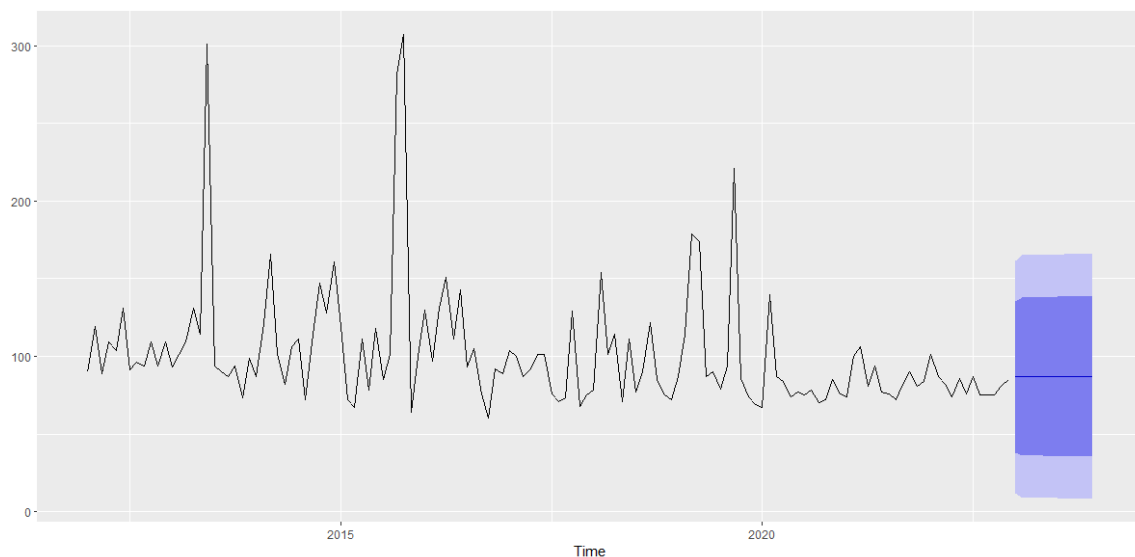

Figure 4. Forecast of API in Shah Alam from January 2023 to December 2023 with ARIMA(1,0,1)

Table 5. Forecasts of API Values for 2023 using ARIMA (1,0,1)

Month	Forecast	95% Confidence Interval
January	82.5010	(43.0386, 121.9634)
February	81.7968	(40.5118, 123.0818)
March	81.6648	(40.0629, 123.2667)
April	81.6400	(39.9345, 123.3454)
May	81.6354	(39.8692, 123.4016)
June	81.6345	(39.8171, 123.4519)
July	81.6344	(39.7680, 123.5008)
August	81.6343	(39.7195, 123.5491)
September	81.6343	(39.6713, 123.5973)
October	81.6343	(39.6231, 123.6455)
November	81.6343	(39.5750, 123.6936)
December	81.6343	(39.5269, 123.7416)

Figure 4 depicts an indication of graph in which the positive outcomes will be under control for the next 12 months or until December 2023. The horizontal axis stands for the number of months from January 2012 onward. The graph also showcases the forecasted monthly Air Pollution Index (API) based on evaluated data up to December 2022. Table 5 displays the specific forecasted values derived from ARIMA (1,0,1). The trend line indicates a consistent decrease, averaging 0.025 percent, suggesting a progressive decline in the API value each month until December 2023. The results imply that the API readings from January to December 2023 will fall within the moderate (51-100) air quality index range by referring to Table 1. Furthermore, all forecasted API values are within 95% confidence intervals.

5. Conclusion

Five ARIMA models were proposed based on the simplest model by estimation value: ARIMA (2,0,2), ARIMA (1,0,2), ARIMA (1,0,1), ARIMA (1,0,3), and ARIMA (0,1,2). Through observation, ARIMA (1,0,1) was found to have the smallest values of AIC and BIC, making it the best model for the Box-Jenkins Methodology. Additionally, Holt's method was identified as the best model for Univariate Techniques, as it exhibited the lowest evaluation values of RMSE and MAE compared to Double Exponential Smoothing. The first goal of the study was to evaluate the best model among the Box-Jenkins Methodology and Univariate Techniques. The results showed that ARIMA (1,0,1) had the lowest value across all error measurements.

The second goal of this study was to forecast the Air Pollution Index in Shah Alam one year ahead, from January 2023 to December 2023 by using the best model. The forecasting was done by using ARIMA (1,0,1) model, and the results revealed that the trend line of anticipated values remained constant at a range of 81.63 and 82.50, which indicates a satisfactory category because the readings fell within the moderate reading (51-100) of API status (Table 1).

For future research, it is recommended to examine how deep learning, or the combination of neural networks and data mining approaches, could enhance the effectiveness of forecasting models. Creating hybrid models that integrate the strengths of different approaches could lead to more accurate forecasts of the air quality index in Shah Alam [24].

Conflict of Interest
The authors declare no conflict of interest in the subject matter or materials discussed in this manuscript.

References

[1]   K. Morrissey *et al.*, "The effects of air quality on hospital admissions for chronic respiratory diseases in Petaling Jaya, Malaysia, 2013–2015," *Atmosphere*, vol. 12, no. 8, p. 1060, Aug. 2021. doi:10.3390/atmos12081060.

[2]   S. Subramaniam *et al.*, "Artificial intelligence technologies for forecasting air pollution and human health: A narrative review," *Sustainability*, vol. 14, no. 16, pp. 9951, Aug. 2022, doi:10.3390/su14169951.

[3]   Department of Environment Malaysia (DOE), *A Guide to Air Pollutant Index (API) in Malaysia*. Kuala Lumpur, Malaysia: Department of Environment, 2000.

[4]   J. Shen, D. Valagolam, and S. McCalla, "Prophet forecasting model: A machine learning approach to predict the concentration of air pollutants (PM2.5, PM10, O3, NO2, SO2, CO) in Seoul, South Korea," *PeerJ.*, vol. 8, pp. e9961, Sep. 2020, doi: 10.7717/peerj.9961.

[5]   W. N. W. M. Din and N. Shaadan, "Application of functional time series model in forecasting monthly diurnal maximum API curves: A comparison between multi- step ahead and iterative one-step ahead approach," *Mal. J. Fund. Appl. Sci.*, vol. 18, pp. 124-137, 2022.

[6]   N. Shaadan and W. N. W. M. Din, "Application of functional time series model in forecasting monthly diurnal API curves: A comparison between multi- step ahead and iterative one-step ahead approach," *Mal. J. Fund. Appl. Sci.*, vol. 18, no. 1, pp. 124-137, Feb. 2022, doi: 10.11113/mjfas.v18n1.2435.

[7]   K. S. Wong, Y. J. Chew, S. Y. Ooi, and Y. H. Pang, "Toward forecasting future day air pollutant index in Malaysia," *The Journal of Supercomputing*, vol. 77, no. 5, pp. 4813–4830, May 2021, doi:10.1007/s11227-020-03463-z.

[8]   C. B. Pande *et al.*, "Daily scale air quality index forecasting using bidirectional recurrent neural networks: Case study of Delhi, India," *Environ. Pollut.*, vol. 351, p. 124040, Jun. 2024, doi: 10.1016/j.envpol.2024.124040.

[9]   R. Das, A. I. Middya, and S. Roy, "High granular and short-term time series forecasting of PM 2.5 air pollutant - a comparative review," *Artif. Intell. Rev.*, vol. 55, no. 2, pp. 1253-1287, Feb. 2022, doi: 10.1007/s10462-021-09991-1.

[10]   J. K. Sethi and M. Mittal, "Analysis of air quality using univariate and multivariate time series models," in *Proc. of the Confluence 2020 -10th Int. Conf. on Cloud Computing, Data Science and Engineering*, 2020, doi: 10.1109/confluence47617.2020.9058303.

[11]   V. Naveen and N. Anu, "Time series analysis to forecast air quality indices in Thiruvananthapuram District, Kerala, India," *Int. J. Eng. Res. Appl.*, vol. 07, no. 06, pp. 66-84, Jun. 2017, doi: 10.9790/9622-0706036684.

[12]   B. A. A. Abdulali and N. Masseran, "Artificial Neural Network (ANN) and ARIMA models for better forecast of the air pollution data in Malaysia," *Sch. J. Phys. Math. Stat.*, vol. 8, no. 10, pp. 184–196, Dec. 2021, doi:10.36347/sjpms.2021.v08i10.001.

[13]   K. Kumar, A. K. Yadav, M. P. Singh, H. Hassan, and V. K. Jain, "Forecasting daily maximum surface ozone concentrations in Brunei Darussalam—an ARIMA Modeling approach," *J. Air Waste Manag. Assoc.*, vol. 54, no. 7, pp. 809-814, Jul. 2004, doi: 10.1080/10473289.2004.10470949.

[14]   J. Zhu, "Comparison of ARIMA model and exponential smoothing model on 2014 air quality index in Yanqing County, Beijing, China," *Appl. Comput. Math.*, vol. 4, no. 6, p. 456, Jan. 2015, doi: 10.11648/j.acm.20150406.19.

[15]   M. Ansari and M. Alam, "An intelligent IoT-Cloud-Based air pollution forecasting model using univariate Time-Series analysis," *Arab. J. Sci. Eng.,* vol. 49, no. 3, pp. 3135–3162, May 2023, doi: 10.1007/s13369-023-07876-9.

[16]   S. Nurman, M. Nusrang, and N. Sudarmin, "Analysis of rice production forecast in Maros District using the Box-Jenkins method with the ARIMA model," *ARRUS J. Math. App. Sci.*, vol. 2, no. 1, pp. 36-48, Feb. 2022, doi: 10.35877/mathscience731.

[17]   R. Thapa, S. Devkota, S. Subedi, and B. Jamshidi, "Forecasting area, production and productivity of vegetable crops in Nepal using the Box-Jenkins ARIMA model," *Turk. J. Agric. - Food Sci. Technol.*, vol. 10, no. 2, 2022, doi: 10.24925/turjaf.v10i2.174-181.4618.

[18]   M. A. A. Bakar, N. M. Ariff, M. S. M. Nadzir, O. L. Wen, and F. N. A. Suris, "Prediction of multivariate air quality time series data using Long Short-Term Memory Network," *Mal. J. Fund. Appl. Sci.*, vol. 18, no. 1, pp. 52–59, Feb. 2022, doi:10.11113/mjfas.v18n1.2393.

[19]   Q. Zhou, Z. Chen, Z. Cai, and Z. Xia, "Prediction of the best portfolio for Bitcoin and gold based on the ARIMA model," *Front. Bus. Econ. Manag.*, vol. 4, no. 3, Aug. 2022, doi: 10.54097/fbem.v4i3.1284.

[20]   D. Adedia, S. Nanga, S. K. Appiah, A. Lotsi, and D. A. Abaye, "Box-Jenkins' methodology in predicting maternal mortality records from a public health facility in Ghana," *Open J. Appl Sci.*, vol. 08, no. 06, pp. 189-202, Jan. 2018, doi: 10.4236/ojapps.2018.86016.

[21]   M. A. Lazim, *Introductory Business Forecasting: A Practical Approach*. Shah Alam, Selangor: UiTM Press, 2011.

[22]   N. L. A. Rani, A. Azid, S. I. Khalit, H. Juahir, and M. S. Samsudin, "Air pollution index trend analysis in Malaysia, 2010-15," *Pol. J. Environ. Stud.*, vol. 27, no. 2, pp. 801–807, Jan. 2018, doi:10.15244/pjoes/75964.

[23]   O. L. H. Leh, S. Ahmad, K. Aiyub, Y. M. Jani, and T. K. Hwa, "Urban air environmental health indicators for Kuala Lumpur city," *Sains Malaysiana*, vol. 41, no. 2, pp. 179-191, 2012.

[24]   J. W. Koo, S. W. Wong, G. Selvachandran, H. V. Long, and L. H. Son, "Prediction of air pollution index in Kuala Lumpur using fuzzy time series and statistical models," *Air Qual. Atmos. Health*, vol. 13, no. 1, pp. 77–88, 2020, doi:10.1007/s11869-019-00772-y.

Biography of all authors

Picture	Biography	Authorship contribution
	Nur Hafiraniza binti Bakhtiar is a full-time student studying for a Bachelor of Science (Hons.) in Statistics at UiTM Seremban Campus and graduated in 2023. Her expertise includes Statistical Modelling and Forecasting.	Data analysis and thesis writing
	Isnewati Ab Malek is a Senior Lecturer at Universiti Teknologi MARA (UiTM), Negeri Sembilan Branch, Seremban Campus. Her area of specialization includes time series analysis and statistical modelling.	Writing – review and editing, supervision
	Haslinda Ab Malek is a Senior Lecturer at Universiti Teknologi MARA (UiTM), Negeri Sembilan Branch, Seremban Campus. Her expertise lies in Environmental Statistics and Statistical Modelling.	Writing – review and editing, validation
	Siti Sarah Januri is a Senior Lecturer at Universiti Teknologi MARA (UiTM), Negeri Sembilan Branch, Seremban Campus. Her expertise lies in Operations Research and Stochastic Modelling.	Review and editing
	Jaida Najihah Jamidin is a lecturer at Universiti Teknologi MARA (UiTM), Negeri Sembilan Branch, Seremban Campus. Her experties lies in Regression Analysis and Statistical Modelling.	Review and formatting

Reliability and Construct Validity Assessment of the Student Competency Questionnaire among Final Year Diploma Students: A Statistical Analysis Approach

Nor Faezah Mohamad Razi
Faculty of Computing,Informatics and Mathematics, Universiti Teknologi MARA, Perak Branch Tapah Campus, Perak, Malaysia
norfa122@uitm.edu.my

Norhayati Baharun
Faculty of Computing,Informatics and Mathematics, Universiti Teknologi MARA, Perak Branch Tapah Campus, Perak, Malaysia
norha603@uitm.edu.my

ABSTRACT

The university strives to provide students with a curriculum that is pertinent and taught by educators with exceptional delivery methods, in a stimulating and dynamic learning setting. The goal is for students to graduate as individuals who can make positive contributions to society, start their own businesses, and take on leadership roles in their professions. This research aims to develop the Student Competency Questionnaire (SCQ) and assess its reliability and validity. Using classical test theory, the evaluation focused on reliability, construct validity, and content validity. A postal questionnaire containing 49 items was distributed in October 2022 to a sample of 59 final year diploma students from various programs. Three experts in education, statistical modeling, and decision making evaluated the content validity of the scale. The Content Validity Index (CVI) for all constructs ranged from 0.96 to 1, exceeding the threshold of 0.70, indicating that the items are 'content valid.' Cronbach's alpha values showed a high level of internal consistency, all exceeding 0.90. Additionally, all items within each construct were highly related to one another, with correlations exceeding 0.30.This research demonstrates that the Pillar 1 education 5.0 @ Universiti Teknologi MARA (UiTM) frameworks effectively capture essential skills for students' academic and professional development, enhancing their employability. The focus on Personal, Adaptive, Digital, 21st Century, and Social competencies, along with student satisfaction, ensures students are equipped with diverse skills needed to thrive in dynamic environments. Overall, UiTM's approach prepares students for success in their future careers and personal lives.

Corresponding Author:
Nor Faezah Mohamad Razi
School of Mathematical Sciences,College of Computing,Informatics and Mathematics
UiTM Perak Branch,Tapah Campus
Email: norfa122@uitm.edu.my

1. Introduction

Academic accomplishment is a crucial indicator for evaluating a student's learning condition and educational advancement. It acts as a thorough measure, covering not just the gaining of knowledge but also the use of critical thinking skills, problem-solving capabilities, and a profound comprehension of the topic. These accomplishments go beyond just grades or test results; they represent the culmination of a student's commitment, hard work, and involvement in the learning process.

During the learning process, a comprehensive evaluation of a student's intellectual development involves showcasing their ability to absorb material, engage actively in class, and

overcome the obstacles encountered in their academic path. Therefore, it plays a vital role in influencing the storyline of a student's educational journey and future opportunities. Educational administration must assess students' competency to ensure their overall success.

Study by [1] have compiled many methods for measuring student accomplishment in higher education, including examinations, tests, grades, GPA, projects, assessments, class participation, portfolios, and peer assessments. Nowadays, the education system has adopted a fresh method to equip pupils to tackle the evolving environment. An intuitive approach is necessary to build and develop a fluid, dynamic, and organic curriculum to ensure students are well-prepared for the labor market.

At UiTM, the Education 5.0@UiTM initiative has developed a framework with five primary pillars that can be applied to the curriculum, learning experience, learning environment, instructors, and delivery. Students will engage with a pertinent curriculum, guided by instructors with excellent teaching methods, in a stimulating and dynamic learning environment. When students graduate from university, they will have had a fulfilling journey that has prepared them to be valuable members of society, entrepreneurs, and capable leaders who can succeed in the professional realm.

Therefore, emphasizing Pillar 1, which involves a cohesive and relevant curriculum, is essential to ensure that students are well-prepared. Five primary abilities focused on for students in the 21st century are social competence, adaptability, digital proficiency, and high levels of personal competency. This research aims to assess student skills by gathering self-evaluation data using a questionnaire.

Questionnaires are vital instruments in survey research, but researchers must recognize and understand their inherent limits. Having this insight is crucial for creating surveys that yield precise and significant outcomes. Researchers can address possible concerns by acknowledging these limits, allowing them to make educated judgements when designing and analyzing survey surveys. Integrating questionnaires with other research methodologies or validation approaches might enhance the overall strength of a study.

In this sense, this study had developed a Student Competencies Questionnaire (SCQ) based on Education 5.0@UiTMI framework (Pillar 1). The SCQ is based mainly on Likert Scale (ranging from 1(less competency) to 10 (most competent) to measure each of the theoretical constructs as presented below. Since framework still not been evaluated in any previous study, therefore the main objective in this study is to evaluate the validity and reliability of SCQ so that able to allows researchers to develop a more well-rounded and reliable investigation that leverages the strengths of diverse research methods and validation techniques. This comprehensive approach enhances confidence in the study's findings and contributes to a deeper understanding in this study. The key to research process in this study been illustrated in Figure 1.
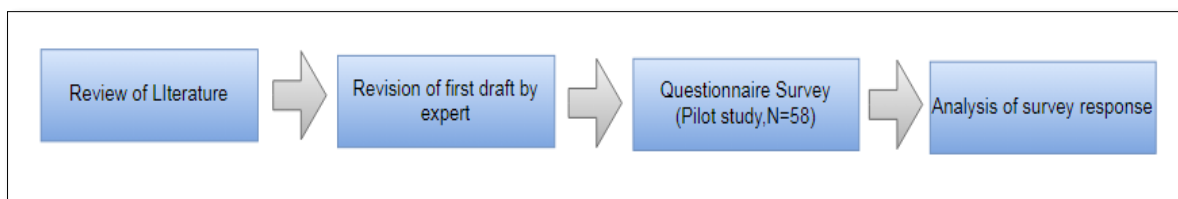

Figure 1: The Summary of Questionnaire Development and Validation

2. Literature Review

2.1 Education 5.0@UiTM: Navigating Pillar 1 framework

Industrial Revolution 4.0 (IR4.0) provides a new catalyst for the change of the current education system in Malaysia. It is driven by technological advancements such as artificial intelligence, virtual reality, data analytics, and the Internet of Things. Based on key findings outlined by [2], [3] present graduates lacked understanding of the IR4.0 concept; they are unprepared for future careers because they relied too heavily on their academic programs to train them and universities are not adequately preparing students for jobs. Therefore, universities have to regularly review the relevancy of their current academic programs especially to prepare students with necessary skills for future IR4.0 workforce. In fact, according to [4],besides knowledge and technical skills, universities also have to equip their students with extra soft skills to enhance their critical

thinking, problem- solving, leadership skills, and lifelong learning to fulfill the changing demands of the IR4.0 job market.

Universiti Teknologi MARA (UiTM), the largest public university in Malaysia, is consistently ranked as the most famous place to study. In the past 60 years, the university has grown from an institution to a large university that manages academic matters of 27 faculties and 4 academic centres comprising 526 programmes at the main campus and 35 state, branch and satellite campuses nationwide. Since 2016, the university has actively engaged in enhancing existing academic programs, launching data analytics labs and smart classrooms in various faculties and state campuses to support the country's IR4.0 initiative.

Apart from this, UiTM has recently launched a new brand of academic ecosystem named Education 5.0@UiTM intending to humanize higher education learning in response to IR4.0 (Refer Figure 2). Education 5.0@UiTM is defined as a learning-centric ecosystem that is sustainable, balanced and principled, driven by values and concepts of Adab and Amanah, powered by intellect and afforded by new, ubiquitous technologies [5]
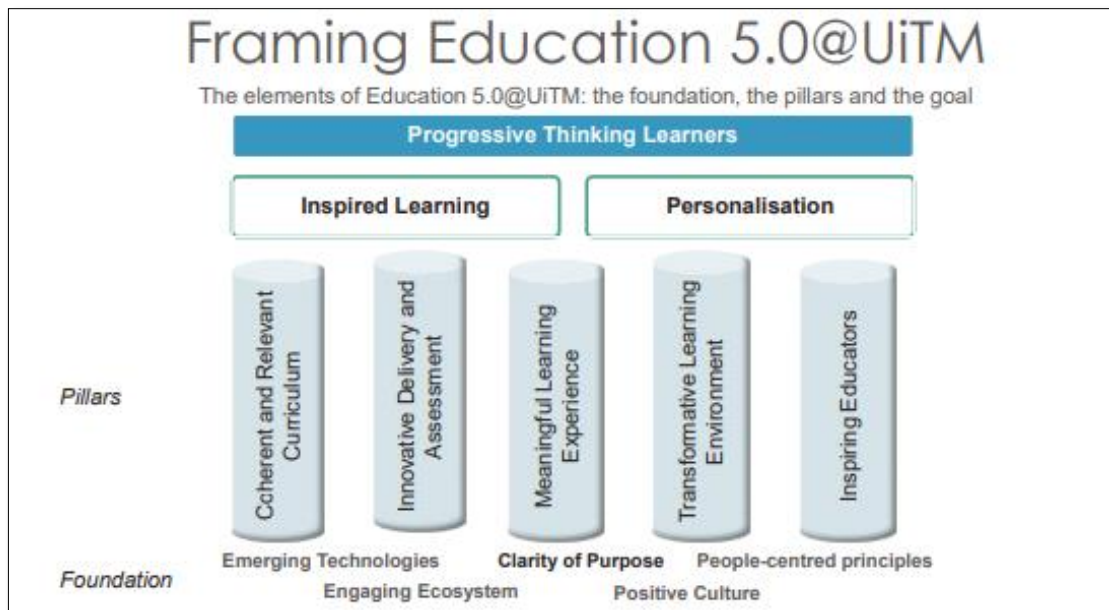
Figure 2: Framing Education 5.0 @UiTM

The Education 5.0 framework at UiTM comprises three key elements: the goal, the pillars, and the foundation. The overarching objective is to cultivate forward-thinking learners characterized by creativity, innovation, and adaptability, enabling them to become versatile professionals, job creators, and leaders in the future. This goal is achieved through the implementation of five pillars: curriculum, delivery, learning experience, learning environment, and educators.

These pillars are guided by a clear sense of purpose, a positive culture, the incorporation of relevant emerging technologies, an engaging ecosystem, and principles centered on people. By aligning these elements, the framework aims to inspire learning and personalization, ultimately nurturing individuals who are well-equipped for the challenges of the future. Thus, to make sure the students embarking into a workforce world with preparedness, this framework of Pillar 1: Coherent and Relevant Curriculum as illustrated in Figure 3.
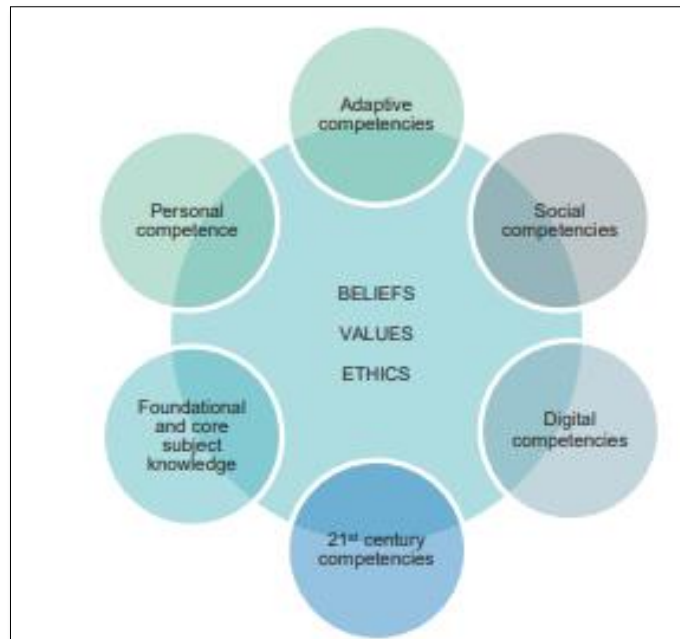
Figure 3: Pillar 1-Coherent and Relevant Curriculum

Students must be holistically prepared to face the world. Holistic preparation means that students should achieve a well-rounded excellence, not just focusing solely on academic curriculum. Therefore, an intuitive approach is necessary for designing and developing a coherent and relevant curriculum. Each program offered should encompass the overall excellence of students, addressing both academic and competency enhancements in every area. Emphasis on these six competencies has been incorporated into Education 5.0@UiTM to ensure that student excellence is achieved comprehensively.

2.2 The connection between six main competencies in Pillar 1 with academic Achievement Efficiency

Competencies are referring to ability to generally understand and perform anything at a basic level. Competence also refers to your ability to comprehend actions or knowledge throughout different parts of life. This means competence occurs at various stages of life as you grow older, learn new things, meet new people and experience new environments. In education field, student competencies are pivotal as they extend beyond mere academic knowledge, encompassing a spectrum of skills, attitudes, and behaviors crucial for overall personal and academic triumph.

The significance of student competencies lies in their capacity to prepare students for success across diverse life domains, transcending academic achievements. In the contemporary workforce, skills like critical thinking, creativity, communication, collaboration, and adaptability are increasingly esteemed. The essence of student competencies in education lies in furnishing a framework that embraces a comprehensive and forward-looking approach to learning. By concentrating on competencies, education not only equips students for academic success but also for a prosperous and gratifying life in the face of rapid societal changes.

Past research has brought attention to the significance of directing focus towards a range of student competencies. These encompass digital competencies as explored by [6] which are essential for navigating the digital landscape. Additionally, the emphasis on 21st-century competencies, as investigated by [7] underscores the importance of skills like critical thinking, creativity, and collaboration in the modern era. Social competency been explored by [8] highlighted the interpersonal skills crucial for effective communication and collaboration. [9] have delved into personal competency, shedding light on the importance of individual attributes and self-awareness.

Furthermore, the research conducted by [10] has contributed insights into adaptive competency, emphasizing the ability to navigate and thrive in dynamic and changing environments. Collectively, these studies underscore the multidimensional nature of student competencies crucial for their comprehensive development. The belief is that with these competencies, students are well-equipped to succeed comprehensively.

In this study, a framework has been developed to examine the interconnection between student competencies, academic achievement, and the satisfaction derived from pursuing a diploma.. However, the focus of this study is specifically limited to evaluating satisfaction with university services as a key element influencing overall student achievement as depicted in Figure 4.
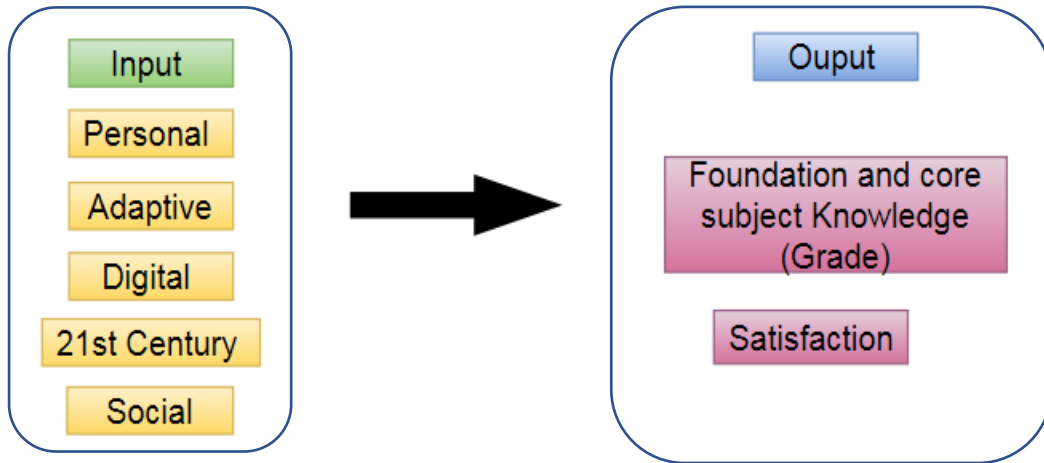

Figure 4: Conceptual model for determining Student academic achievement efficiency

3. Methodology

3.1 The Development of Student Competency Questionnaire (SCQ)

In this paper, Student's Competency Questionnaire been developed consisting of four parts. Part A focuses on the demographics of students where seven items been asked. Part B on Academic Background involving the history of student achievement during school day on selected four selected subjects. Meanwhile for Part C focused on Student Competencies factors as stated in Pillar 1 education 5.0@uitm.Part D on Student Satisfaction. Table 1 shows the description of the questionnaire for this study.

The construction of the questionnaire is based on the adaptation from previous studies. Whenever possible, items for Part C have been built as 10 Likert scales ranging from Not Competence at all to Very competent. This study prefers to use 10 Likert scale compares to 7 Likert scale like commonly social study used to provide more granularity by allowing for a finer differentiation of response. In fact a study by [11] revealed that 10 points of Likert scale is more efficient than 5 points of Likert scale in operating of measurement model.

3.2 Testing SCQ through Pilot Study

A pilot study provides an excellent opportunity to uncover such problems ahead of time, minimizing the need to adapt procedures or to develop contingency plans on short notice when the larger study is being conducted. Thus, this study was conducted to assess the feasibility and effectiveness of the research design and data collection methods planned in this study. This small-scale investigation aimed to identify potential challenges, refine the research instruments, and ensure the smooth implementation of the larger study. The pilot study employed a quantitative approach, utilizing a cross-sectional design. The key variables under investigation included five competencies factor and one satisfaction. Before the process of collection data started, an application to conduct the study will be made to Research Ethic Committee, Program Coordinator and Student Affair officer in order inform and get permission regarding this study. Participants were selected using a convenience sampling method, with inclusion criteria (final year diploma students). A total of 59 final year diploma students were recruited for the pilot study considering with 95% confidence and 5% probability the practical constraints of time and resources.

Table 1: The description of Student's Competency Questionnaire

Component	Reference
Part A:Demographic Profile	
Gender Semester Hometown Family Income Education Aids Mother's Level of Education Father's Level of Education	-
Part B:Academic Background	
SPM Result: Mathematic Additional Mathematic Science English Student Grade(CGPA) Student Grade (GPA)	- - - - - - -
Part C:Competency Factors	
Personal Competencies	Adapted from [12]
Adaptive Competencies	Adapted from [13]
Digital Competencies	Adapted from [14]
21st Century Competencies	Adapted from [7]
Social Competencies	Adapted from [15]
Part D: Other Factors	
Student Satisfaction	Adapted from [16], [17]

3.3 The Validation Process of Student Competency Questionnaire (SCQ)

According to [18] content validity is the degree to which a measurement reveals the specific intended domain of content. Content validity, also identified as logical or rational validity, is the assessment of how much the item measure or represents the construct of the research. Earlier, [19] states that content validity are the determination of the content representativeness or content relevance of the elements or items of an instrument by the application of two-stage process: the development and judgment. The CVI is derived from the content relevance of the items based on the four-point scale rating given by the experts. The actual CVI is the proportion of items that obtained a rating of 3 or 4 by the experts divide with the total items in the construct [20].Established on [21] suggestion, the content validity index (CVI) must be at least 0.70 to show higher content validity.

To obtain expert reviews have become a widespread practice in questionnaire development [22] .Researchers have often requested subject matter experts, methodology experts, and language experts to appraise the draft questionnaires and provide a critique on the items as a technique of recognizing questionnaire problems, potential measurement errors or a breakdown in the question answering process. To determine the number of expert reviewers for content validity of the questionnaire, These suggest that the number of expert reviewers for content validity should between five to 20 [23], [24].However, [19] argues that a minimum of three experts should be employed but not to exceed ten experts.

Thus, this study decided to appoint three expert reviews (PT, DN and DF) where each of these experts is an academician and PhD holder that have expertise in their field. Expert 1 namely as PT expertise in Educational Technology, Expert 2 namely as DN expertise in Statistical Modeling while Expert 3 namely as DF expertise in on Decision Making area. Content validity was attended to ensure that the items in the questionnaire had been adopted and customized adequately. The review of the items is carried out by five experts; three experts in the content of the subject, one expert in the field of methodology, and one language expert. The experts were requested to value each item on the relevance of the construct according to Table 2 below:

Table 2: The ranking of each constructs/items in SCQ

Statement	Score
Not Relevant (Requires Major amendment)-NA	1
Somewhat relevant (Requires minor amendment-BE	2
Relevant (No amendment but need minor revision)-IA	3
Highly relevant (No amendment)- EE	4

The content validity approach can be conducted through the face to face approach or non-face to face approach. This study decided to use non face to face approach (online) for Expert 1 due to long distance reason while Expert 2 and Expert 3 were conducted through face to face approach. After the discussions with all the content experts, the researcher builds a summary of all responses and comments. From there, the researcher calculates the 'content validity index (CVI)', the most widely used quantification of content validity.

3.4 Items Internal Consistency

Internal consistency is a critical measure for assessing whether scale items effectively capture the intended construct. Commonly used indices for this purpose include Cronbach's alpha, corrected item-total correlation, and average inter-item correlation [25]–[27] .Cronbach's alpha is widely utilized to gauge the extent to which items in a multi-item scale measure the same underlying concept, while corrected item-total correlation assesses an item's consistency with the overall test score. Researchers in [28] suggested reliability thresholds, as presented in Table 3.

Table 3. Interpretation of Cronbach alpha-score

Cronbach alpha score	Interpretation
<0.5	Items need to be dropped
<0.6	Items need to be dropped
0.6-0.7	Acceptable
0.7-0.8	Good and acceptable
0.8-1.0	Very good and effective level of consistency

Inter-item correlations reveal how items within a scale are interrelated, aiding in identifying item redundancy and relatedness. Researchers in [25], [29] recommended that both the average inter-item correlation and the average inter-item correlation for each item should fall within the range of 0.15–0.50, with a corrected item-total correlation ideally exceeding 0.30 [30] Items failing to meet these criteria were excluded from the scale.

4. Results and Discussion

4.1 Content Validity

The CVI for all the constructs of this study ranged from 0.96 to 1 (Table 4) found to be above the threshold of 0.70 to demonstrate that the items in each of the constructs are approximately 'content valid'.

Table 4: The Result of Content Validity

SCQ content meaningful items	Expert 1	Expert 2	Expert 3
Items rated 'not relevant' and 'somewhat relevant'	2	0	0
Items rated 'relevant' and 'highly relevant'	47	49	49
Total Initial Items	49	49	49
Content Validity Index	0.96	1	1

4.2 Reliability (Internal Consistency)

The results in Tables 5 display the corrected item-total correlations and Cronbach's alpha coefficients for all items across six main constructs: Personal Competency, Adaptive Competency, Digital Competency, 21st Century Competency, Social Competency, and Student Satisfaction. The Cronbach's alpha values indicate a very good level of internal consistency, all exceeding 0.90. Additionally, Pearson Product Moment of correlation was employed to assess the corrected item-total correlations within each construct. The result revealed that for the Personal Competency domain, all item-total correlations were above 0.591 (range: 0.591 to 0.800), for Adaptive Competency, above 0.540 (range: 0.540 to 0.829 for Digital Competency, above 0.629 (range: 0.629 to 0.835), for 21st Century Competency, above 0.717 (range: 0.717 to 0.834), for Social Competency, above 0.723 (range: 0.723 to 0.884), and for Satisfaction, above 0.884. These results indicate that each item within a scale correlates well with the other items in the scale, contributing to the overall score. Overall, all items within each construct are highly related to one another within their respective domains, with correlations exceeding 0.30.

Table 5: Internal Consistency of Personal Competency domain items (N=58)

Domain	Code	Item	Corrected item-total correlation	Cronbach's Alpha
Personal	P1	Self-Control	0.591	0.911
	P2	Trustworthiness	0.735	
	P3	Conscientiousness	0.746	
	P4	Adaptability	0.692	
	P5	Innovativeness	0.633	
	P6	Achievement drive	0.669	
	P7	Commitment	0.701	
	P8	Initiative	0.800	
	P9	Optimism	0.722	
Adaptive	A1	Recognizing skills	0.740	0.924
	A2	Respecting	0.540	
	A3	Capable of making one's own decision	0.808	
	A4	Assuming responsibility	0.744	
	A5	Exercising leadership	0.792	
	A6	Showing interest in the topic	0.783	
	A7	Showing the interest in learning process	0.829	
	A8	Having a flexible attitude	0.729	
Digital	D1	Information Literacy	0.776	0.910
	D2	Computer Literacy	0.795	
	D3	Media Literacy	0.835	
	D4	Communication Literacy	0.776	
	D5	Visual Literacy	0.629	
	D6	Technological Literacy	0.683	
21st Century Competency	21ST 1	Critical Thinking	0.720	0.914
	21ST 2	Collaboration	0.752	
	21ST 3	Communication	0.803	
	21ST 4	Creativity	0.770	
Social	S1	Empathy	0.770	0.924
	S2	Leveraging diversity	0.717	

	S3	Political Awareness	0.797	
	S4	Communication	0.811	
	S5	Conflict management	0.775	
	S6	Team Capabilities	0.834	
Satisfaction	SAT1	the University	0.853	0.971
	SAT2	the Administrative and Student Services	0.884	
	SAT3	the Cafeteria	0.781	
	SAT4	the Facility Condition provided	0.831	
	SAT5	the Library	0.723	
	SAT6	the Class Placements	0.823	
	SAT7	the Lecturers	0.828	
	SAT8	the University Buildings	0.853	
	SAT9	the Relevance of Teaching to Practice	0.834	
	SAT10	the Support from Lecturers	0.824	
	SAT11	the Presentation of Information	0.779	
	SAT12	the Atmosphere among Students	0.844	
	SAT13	the Courses	0.835	
	SAT14	the Reputation of the University	0.873	
	SAT15	the Additional Courses Offered	0.842	
	SAT16	the University Life Outside the Courses	0.752	

5. Conclusion

The present study aimed to validate the competency properties based on the Pillar 1 education 5.0 @ UiTM frameworks, which comprises five main competency domains and student satisfaction for final year diploma students at UiTM. The selected scales for Personal, Adaptive, Digital, 21st Century, and Social competencies were deemed content valid (greater than 0.70) through expert review. Cronbach's alpha values for all domains exceeded 0.90, confirming the adequacy of internal consistencies. Consequently, these findings support the conclusion that the competencies outlined in the Pillar 1 education 5.0 @ UiTM framework are well-represented and reliably measured among final year diploma students at UiTM.

The validation of competencies among final year diploma students at UiTM has significant implications. The confirmed content validity and high internal consistency of the competency domains indicate that the Pillar 1 education 5.0 @ UiTM framework effectively captures essential skills for students' academic and professional development, enhancing their employability. The focus on Personal, Adaptive, Digital, 21st Century, and Social competencies, along with student satisfaction, ensures students are equipped with diverse skills needed to thrive in dynamic environments. These findings suggest that UiTM is preparing students holistically, fostering their success in both future careers and personal lives.

Conflict of Interest
The authors declare no conflict of interest in the subject matter or materials discussed in this manuscript.

References

[1] J. Márquez, L. Lazcano, C. Bada, and J. L. Arroyo-Barrigüete, "Class participation and feedback as enablers of student academic performance," *SAGE Open*, vol. 13, no. 2, p. 21582440231177298, Apr. 2023, doi: 10.1177/21582440231177298.

[2] A. N. Azmi, Y. Kamin, M. K. Noordin, and A. N. M. Nasir, "Towards industrial revolution 4.0: Employers' expectations on fresh engineering graduates," *Int. J. Eng. Technol.*, vol. 7, no. 4.28, pp. 267–272, Dec. 2018, doi: 10.14419/ijet.v7i4.28.22593.

[3] M. Z. Kamde, S. Mamter, M. E. Mamat, and N. A. Abdullah@Dolah, "Employer's expectation to QS graduates in facing industrial revolution 4.0," in *Virtual Go Green: Conference and Publication (v-GoGreen 2021)*, Sep. 29-30, 2021, pp. 80–86. [Online]. Available: https://ir.uitm.edu.my/id/eprint/65650

[4] M. I. M. Salleh, Z. Ibrahim, K. I. Othman, H. Yusoff, and S. Ariffin, "Educator acceptance of Education 5.0@UiTM framework and initiatives: A descriptive analysis," *Int. J. e-Learning High. Educ.*, vol. 12, no. 1, pp. 99–108, Jan. 2020, doi: 10.24191/ijelhe.v12n1.1214.

[5] UiTM, "UiTM academic compass education 5.0@UiTM: Navigating the future," in *UiTM Academic Compass Education 5.0@ UiTM Navigating the Future*. UiTM Press, 2019.

[6] F. H. Ohl, "Perspectives of digital competencies. A comparison of different constructs of digital pedagogical competencies," *Zeitschrift für Hochschulentwicklung*, vol. 19, no. 1, pp. 171-188, Mar. 2024, doi: 10.21240/zfhe/19-01/09.

[7] T. R. Kelley, J. G. Knowles, J. Han, and E. Sung, "Creating a 21st century skills survey instrument for high school students," *Am. J. Educ. Res.*, vol. 7, no. 8, pp. 583–590, Aug. 2019, doi: 10.12691/education-7-8-7.

[8] C. Junge, P. M. Valkenburg, M. Deković, and S. Branje, "The building blocks of social competence: Contributions of the consortium of individual development," *Dev. Cogn. Neurosci.*, vol. 45, p. 100861, Oct. 2020, doi: 10.1016/j.dcn.2020.100861.

[9] P. Martínez-Clares and C. González-Lorente, "Personal and interpersonal competencies of university students entering the workforce: Validation of a scale," *Relieve*, vol. 25, no. 1, pp. 1–18, May 2019, doi: 10.7203/relieve.25.1.13164.

[10] A. S. Bahari, M. I. Mahmud, M. M. Wahed, A. D. Adnan, and N. S. A. Khidir, "Challenges and emotional well-being of students in facing new learning norms," *Int. J. Arts and Soc. Sci.*, vol. 4, no. 3, pp. 194–202, 2021.

[11] Z. Awang, A. Afthanorhan, and M. Mamat, "The Likert scale analysis using parametric based Structural Equation Modeling (SEM)," *Comput. Methods Soc. Sci.*, vol. 4, no. 1, pp. 13–21, Apr. 2016.

[12] B. Dong, X. Peng, and N. Jiang, "Exploring the domain of emotional intelligence in organizations: Bibliometrics, content analyses, framework development, and research agenda, " *Front. Psychol.*, vol. 13, p. 810507, Mar. 2022, doi: 10.3389/fpsyg.2022.810507.

[13] P. Migliorini and G. Lieblein, "Facilitating transformation and competence development in sustainable agriculture university education: An experiential and action oriented approach," *Sustainability*, vol. 8, no. 12, p. 1243, Nov. 2016, doi: 10.3390/su8121243.

[14] P. Reddy, B. Sharma, and K. Chaudhary, "Measuring the digital competency of freshmen at a higher education institute," in *Proc. 23rd Pacific Asia Conf. Inf. Syst. (PACIS 2020)*, Dubai, UAE, Jun. 22, 2020.

[15] E. N. Leganés-Lavall and S. Pérez-Aldeguer, "Social competence in higher education questionnaire (CCSES): Revision and psychometric analysis," *Front. Psychol.*, vol. 7, p. 1484, Oct. 2016, doi: 10.3389/fpsyg.2016.01484.

[16] M. I. Zakaria, M. F. A. Hanid, and R. Hassan, "Validity of instrument to measure mathematics teachers' perceptions towards problem-based learning activities," *Int. J. Eval. & Res. Educ.*, vol. 13, no. 1, pp. 355-362, Feb. 2024, doi: 10.11591/ijere.v13i1.26739.

[17] M. Ikram and H. B. Kenayathulla, "Education quality and student satisfaction nexus using instructional material, support, classroom facilities, equipment and growth: Higher education perspective of Pakistan," *Fron. Educ.*, vol. 8, p. 1140971, Mar. 2023, doi: 10.3389/feduc.2023.1140971.

[18] E. G. Carmines and R. A. Zeller, *Reliability and Validity Assessment. SAGE*, 1979.

[19] M. R. Lynn, "Determination and quantification of content validity," *Nursing Research*, vol. 35, no. 6. pp. 382–386, Nov. 1986, [Online]. Available: https://journals.lww.com/nursingresearchonline/citation/1986/11000/DeterminationandQuantificatonOfContent.17.aspx

[20] C. F. Waltz and B. R. Bausell, *Nursing Research: Design,Statistics and Computer Analysis.* Davis F A, 1981.

[21] E. Almanasreh, R. Moles, and T. F. Chen, "Evaluation of methods used for estimating content validity.," *Res. Soc. Admin. Pharm.*, vol. 15, no. 2, pp. 214–221, Feb. 2019, doi: 10.1016/j.sapharm.2018.03.066.

[22] E. M. Ikart, "Survey questionnaire survey pretesting method: An evaluation of survey questionnaire via expert reviews technique," *Asian J. Soc. Sci. Stud.*, vol. 4, no. 2, p.1, Apr. 2019, doi: 10.20849/ajsss.v4i2.565.

[23] N. Othman, "Questionnaire for teachers (MOQ-T) measurment items using Content Validity Ratio (CVR)," *J. Mech. Continua Math. Sci.*, vol. 15, no. 6, Jun. 2020, doi: 10.26782/jmcms.2020.06.00018.

[24] T. Carli, M. Kosnik, L. Z. Kregeli, G. Burazeri, and A. Kukec, "The APISS questionnaire: A new tool to assess the epidemiology of systemic allergic reactions to bee venom in beekeepers," *Slov. J. Public Heal.*, vol. 62, no. 3, pp. 137–144, Jun. 2023, doi: 10.2478/sjph-2023-0019.

[25] L. A. Clark and D. Watson, "Constructing validity: Basic issues in objective scale development," *Psychol. Assess.,* vol. 7, no. 3, pp. 309-319, 1995. doi: 10.1037/1040-3590.7.3.309.

[26] D. F. Polit and C. T. Beck, *Generating and Assessing Evidence for Nursing Practice*, 9th ed. Philadelphia, PA: Wolters Kluwer, 2015.

[27] F. A. M. Nawi, A. M. A.Tambi, M. F. Samat, and W. M. W. Mustapha, "A review on the internal consistency of a scale: The empirical example of the Influence of human capital investment on Malcom Baldridge Quality Principles in TVET institutions," *Asian People J.*, vol. 3, no. 1, pp. 19–29, Apr. 2020, doi: 10.37231/apj.2020.3.1.121.

[28] I. Kennedy, "Sample size determination in test-retest and Cronbach Alpha reliability estimates," *Br. J. Contemp. Educ.*, vol. 2, no. 1, pp. 17–29, 2022, doi: 10.52589/BJCE-FY266HK9.

[29] L. A. Clark and D. Watson, "Constructing validity: New developments in creating objective measuring instruments," *Psychol. Assess.,* vol. 31, no. 12, p. 1412, Dec. 2019.

[30] J. C. Nunnally and I. H. Bernstein, *Psychometric Theory*, 3rd ed. New York, USA: McGraw-Hill, 1994.

Biography of all authors)

Picture	Biography	Authorship contribution
	Nor Faezah Mohamad Razi is a Senior lecturer in University Teknologi MARA (UiTM). She is the Resource Person for Fundamentals of Research. She received her BSc in Statistics and MSc in Applied Statistics from UiTM in 2009 and 2012 respectively. Currently she pursues her PhD study in Decision Science area. Her research interest in Structural Equation Modeling (SEM), Data Envelopment Analysis and Educational technology. She can be contacted at norfa122@uitm.edu.my.	The research work was designed collaboratively by all authors, with each contributing valuable insights and expertise to the conceptualization of the study. Data collection was primarily carried out by this author who meticulously gathered the necessary information from the final year diploma students at UiTM.
	Dr. Norhayati Baharun is an Associate Professor of Statistics, Universiti Teknologi MARA Perak Branch,Tapah Campus. She received her PhD in Statistics Education from the University of Wollongong Australia in 2012. Her career started as an academic from January 2000 to date at the Universiti Teknologi MARA that specialized in statistics. Other academic qualifications include both Master Degree and Bachelor Degree in Statistics from Universiti Sains Malaysia and Diploma in Statistics from Institute Teknologi MARA. Among her recent academic achievements include twelve on-going and completed	Data analysis and interpretation were conducted jointly, with each author contributing to the statistical analysis and the interpretation of the results. The drafting of the article was a collective effort, with all authors contributing to the writing, reviewing, and editing process to ensure the clarity and accuracy of the final manuscript.

	research grants (local and international), four completed supervision of postgraduate studies, fifteen indexed journal publications, two academic and policy books, twenty-six refereed conference proceedings and book chapter publications, a recipient of 2013 UiTM Academic Award on Teaching, and fourteen innovation projects with two registered Intellectual Property Rights by RIBU, UiTM. She is also a certified Professional Technologist (Ts.) (Information & Computing Technology) of the Malaysia Board of Technologist (MBOT), a Fellow Member of the Royal Statistical Society (RSS), London, United Kingdom, a Professional Member of Association for Computing Machinery (ACM), New York, USA, and a Certified Neuro Linguistic Program (NLP) Coach of the Malaysia Neuro Linguistic Program Academy. Her research interests continue with current postgraduate students under her supervision in the area of decision science now expanding to a machine learning application. She can be contacted at norha603@uitm.edu.my	

E-Learning Application based on Gagne's Instructional Design Model (E-DBLEARN)

Fatin Najwa Kamal
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Terengganu Branch Kuala Terengganu Campus, Terengganu, Malaysia
2021600554@student.uitm.edu.my

Nor Azila Awang Abu Bakar
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Terengganu Branch Kuala Terengganu Campus, Terengganu, Malaysia
azila268@uitm.edu.my

Siti Nurul Hayatie Ishak
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Terengganu Branch Kuala Terengganu Campus, Terengganu, Malaysia
sitinurul @uitm.edu.my

Norlina Mohd Sabri
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Terengganu Branch Kuala Terengganu Campus, Terengganu, Malaysia
norli097@uitm.edu.my

ABSTRACT

Learners often struggle with the complex and interconnected terminology in database environments and data modeling, leading to difficulties in understanding and applying key concepts effectively. In addition, a preliminary study has revealed that inadequate classroom facilities, varying learning styles, and the use of external resources that are not tailored to the syllabus have negatively impacted the understanding of theory-based topics. This study aims to achieve several objectives: identifying requirements of an e-learning application for Introduction to Database, design and develop the e-learning application and evaluating the e-learning application's functionality and usability. This research has been initiated by preliminary analysis on the application requirements based on literature review and surveys. The design phase followed, producing essential diagrams which are navigation maps and storyboards. The development was utilizing Adobe Animate, Adobe Premiere Pro, Adobe Audition and Adobe Photoshop for producing the application for learning database. Lastly, the functionality and usability of the application was verified through testing. Usability testing involved expert evaluation with input from three experts. The results acquired give a positive indicator for further implementation. Further evaluation with 30 respondents from the target users has produced results indicated high satisfaction with the application's overall usability, particularly in visual design and self-assessment and learnability constructs, which achieved the highest mean scores of 4.80 with standard deviations of 0.41 and 0.48, respectively. In conclusion, E-DBLEARN offers university students studying database courses an interactive, multimedia-rich platform expected to boost engagement, understanding, and self-directed learning, while equipping lecturers with an effective tool to teach theoretical concepts in the database courses. Gagne's Instructional Design Model was incorporated to ensure that learning is effective, efficient, and structured by guiding the instructional process through a systematic approach.

Corresponding Author:
Nor Azila Awang Abu Bakar
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Terengganu Branch Kuala Terengganu Campus, Terengganu, Malaysia

Email: azila268@uitm.edu.my

1. Introduction

E-learning, often referred to as online or web-based learning, leverages telecommunication technologies to deliver educational content and training to students through applications or web browsers. This method of learning is reshaping education by offering flexible, accessible, and personalized experiences, making it easier for learners to access information and engage with course material anytime, anywhere [1],[2]. Nowadays, e-learning is increasingly becoming the standard for modern education. Unlike traditional teaching approaches, e-learning allows learners to participate in a structured learning environment from anywhere in the world, enabling the transfer of skills and knowledge to many recipients simultaneously or at different times, leading to greater multidirectional communication [3]. In addition to its accessibility and flexibility, e-learning's effectiveness is further amplified by the integration of multimedia, which enriches the learning environment and supports diverse learning styles.

Multimedia plays a crucial role in e-learning by enhancing the learning experience through the integration of various forms of media such as text, audio, video, graphics, and interactive elements, which together create a more engaging, interactive, and effective educational environment. This combination not only caters to different learning styles but also aids in better retention and understanding of complex concepts, making e-learning a dynamic and versatile approach to education [2],[3]. Multimedia improves e-learning by supporting various learning styles, and its value is especially clear in subjects like database environments, where complex terms can be hard for students to grasp.

Learners often struggle with the complex and interconnected terminology in database environments and data modeling, leading to difficulties in understanding and applying key concepts effectively. A preliminary study was conducted in Universiti Teknologi Mara (Terengganu Branch) focusing on database courses. Numerous database courses, both introductory and advanced, are offered to students across various programs. In the course, students encounter a diverse range of topics that are theoretical, conceptual, and practical, encompassing areas such as data modeling, normalization theories, and hands-on SQL programming [4],[5]. The study has revealed that inadequate classroom facilities, varying learning styles, and the use of external resources that are not tailored to the syllabus have negatively impacted the understanding of theory-based topics.

To address these challenges, it is essential to develop and implement targeted strategies that enhance both the teaching and learning of complex database concepts, ensuring that students are better equipped to grasp and apply the material effectively. Therefore, incorporating interactive multimedia content as additional learning material can be an effective tool to assist students in mastering these topics [2],[3]. The objectives of the study are to identify the requirements of an e-learning application for Introduction to Database course, design and develop the e-learning application and evaluate the e-learning application's functionality and usability. Interactive tools can make learning more engaging and help students see how terms are applied in real scenarios, thereby making the terminology less intimidating and improving overall comprehension [2],[6].

2. Literature Review

2.1 E-Learning Application using Multimedia Technology

E-learning applications have transformed education by providing flexible, accessible learning experiences through digital platforms. These applications offer a range of features such as interactive content, multimedia resources, and assessment tools that cater to diverse learning styles and needs [7]. They support asynchronous learning, allowing students to access materials at their convenience and from various locations, which promotes inclusivity and accommodates different schedules [8]. E-learning also facilitates personalized learning paths through adaptive technologies and data analytics, enhancing the effectiveness of education by tailoring instruction to individual progress and preferences [9]. However, challenges such as the digital divide and varying levels of learner engagement must be addressed to maximize their potential [10].

The integration of multimedia technology within e-learning applications significantly enhances their effectiveness. Multimedia elements, such as videos, animations, and interactive graphics, make learning more engaging by providing visual and auditory stimuli that complement textual information [11]. This multimodal approach caters to different sensory preferences, making complex concepts easier to understand and remember [2]. Moreover, multimedia technology enables the creation of rich, interactive learning environments where students can explore simulations,

27

participate in virtual labs, and experience scenarios that would be difficult to replicate in traditional classrooms [12]. The synergy between e-learning applications and multimedia technology, therefore, not only enriches the learning experience but also helps bridge the gap between theory and practice [13].

2.2 Database Course

Database courses are offered in various programs at multiple levels in the University. A database course is crucial because it equips students with essential skills for managing and structuring data, which is foundational to virtually all areas of computer science. As data-driven decision-making becomes increasingly prevalent across industries, understanding how to efficiently store, retrieve, and manipulate data is vital. The course provides knowledge in key areas such as data modeling, SQL, and database management systems (DBMS), which are critical for designing and maintaining robust and scalable data systems. These skills are not only fundamental for software development but also for emerging fields like data science, artificial intelligence, and cloud computing. Given the centrality of data in today's digital economy, a database course is indispensable for preparing students to tackle complex data challenges and contribute meaningfully to the development of advanced technological solutions.

2.2.1 Database Environment and Data Model

The database environment topic covers the infrastructure, tools, and methodologies used to store, manage, and retrieve data. It includes the database management system (DBMS), which serves as the core software for handling data operations, as well as the hardware, software, users, and procedures that ensure data integrity, security, and availability. The environment supports various tasks such as data definition, manipulation, and control, which are crucial for managing large datasets efficiently in modern applications [14].

Data models, on the other hand, describe the logical structure of a database, defining how data is organized, stored, and interrelated. They serve as blueprints for creating databases and play a vital role in ensuring data consistency and integrity. There are various types of data models, including relational, hierarchical, network, and object-oriented models, each tailored to specific needs and use cases [15]. The relational data model, proposed by E.F. Codd, remains the most prevalent, representing data in tables (relations) that can be queried using SQL. Modern advancements have led to the development of more complex models, such as NoSQL and graph databases, to address the needs of big data and real-time applications [16].

2.3 Gagne Instructional Design Model

The Gagne Instructional Design Model is a systematic approach to instructional design that aims to improve learning outcomes through a structured sequence of events. The model is based on nine instructional events that are designed to guide learners through the process of acquiring new knowledge and skills effectively. The Gagne model contributes to enhancing learning by ensuring that instructional activities are designed to address each critical step in the learning process. It provides a clear framework that helps educators create effective learning experiences that engage learners, support their understanding, and reinforce their skills. The model's structured approach also facilitates the design of assessments and feedback mechanisms that are integral to successful learning outcomes [17]. Table 1 summarizes nine instructional events from Gagne's model.

Table 1. Nine Events in Gagne's Instructional Design Model [17]

Gaining Attention	Informing Learners of Objectives	Stimulating Recall of Prior Learning
Capturing learners' interest to prepare them for learning.	Clearly stating what learners will be able to do after the instruction.	Activating previous knowledge related to the new content.
Presenting the Content	**Providing Learning Guidance**	**Eliciting Performance**
Delivering new information in a structured and organized manner.	Offering hints, examples, and support to help learners understand the content.	Allowing learners to practice and apply what they have learned.
Providing Feedback	**Assessing Performance**	**Enhancing Retention and Transfer**
Offering corrective feedback to reinforce correct performance and address errors.	Evaluating learners' ability to perform the learned skills or knowledge.	Using strategies to help learners retain and apply knowledge in different contexts.

3. Methodology

The development of the E-DBLEARN application was guided by the ADDIE model, a widely recognized instructional design framework that encompasses five phases: Analysis, Design, Development, Implementation, and Evaluation. Each phase played a crucial role in ensuring the systematic creation of an effective e-learning tool tailored to the needs of students and educators in introductory database courses. Table 2 provides a comprehensive overview of the methodology employed for the E-DBLEARN project, detailing the critical steps taken at each phase to ensure the successful development and deployment of the application.

Table 2. E-DBLEARN Project Methodology

Phases	Activities	Outcomes
Analysis	• Distribute questionnaires to target users • Identify problem • Consult with Subject Matter Expert (SME) • Conduct literature review	• User problems • Requirements and content of the e-learning application
Design	• Design navigation map • Design storyboard for e-learning content • Design test plan • Design usability question form for expert	• Navigation map • Storyboard • Test plan • Expert evaluation form
Development	• Develop e-learning application based on designed storyboard	• Completed e-learning application for Introduction to Database (E-DBLEARN)
Implementation	• Test the developed e-learning application functionality	• Published e-learning application. • The results of E-DBLEARN's functionality
Evaluation	• Evaluate the developed e-learning application usability • Conduct expert evaluation	• The results of E-DBLEARN's usability

The first phase, Analysis, focuses on identifying and understanding the needs of the target users, which includes university students and lecturers engaged in database courses. During this phase, questionnaires were distributed, and research was conducted to identify problems in current course delivery for the introduction topics consisting of complex terminologies. Subject Matter Experts (SMEs) were consulted to ensure alignment with educational standards, and an online Google Form questionnaire was used to gather target users' feedback. This phase resulted in identifying key user problems and finalizing the content and functional requirements for the E-DBLEARN application.

In the Design phase, insights from the analysis were translated into a detailed plan for the application's structure and content. A navigation map was created to outline the user interface's layout, ensuring an intuitive experience that facilitates learning. Additionally, a storyboard was developed to visually represent the instructional design, integrating multimedia elements and interactive features for enhanced engagement. The deliverables, including a comprehensive navigation map and storyboard, served as blueprints guiding the application's development. Figure 1 shows the storyboard sample for E-DBLEARN.

Figure 1. Storyboard Sample

The Development phase focuses on constructing the E-DBLEARN application by bringing the design concepts to life, emphasizing the integration of interactive multimedia elements to enhance the learning experience and adapting with Gagne's Nine Event of Instruction Model. Adobe tools were used to develop graphical, animated, video, and audio components, with continuous testing and refinement ensuring a fully functional and engaging e-learning application ready for deployment. Figure 2 illustrates snapshots of the application interface that are developed based on the model selected.

| Present the content | Provide feedback |
| Provide learning guidance | Assess performance |
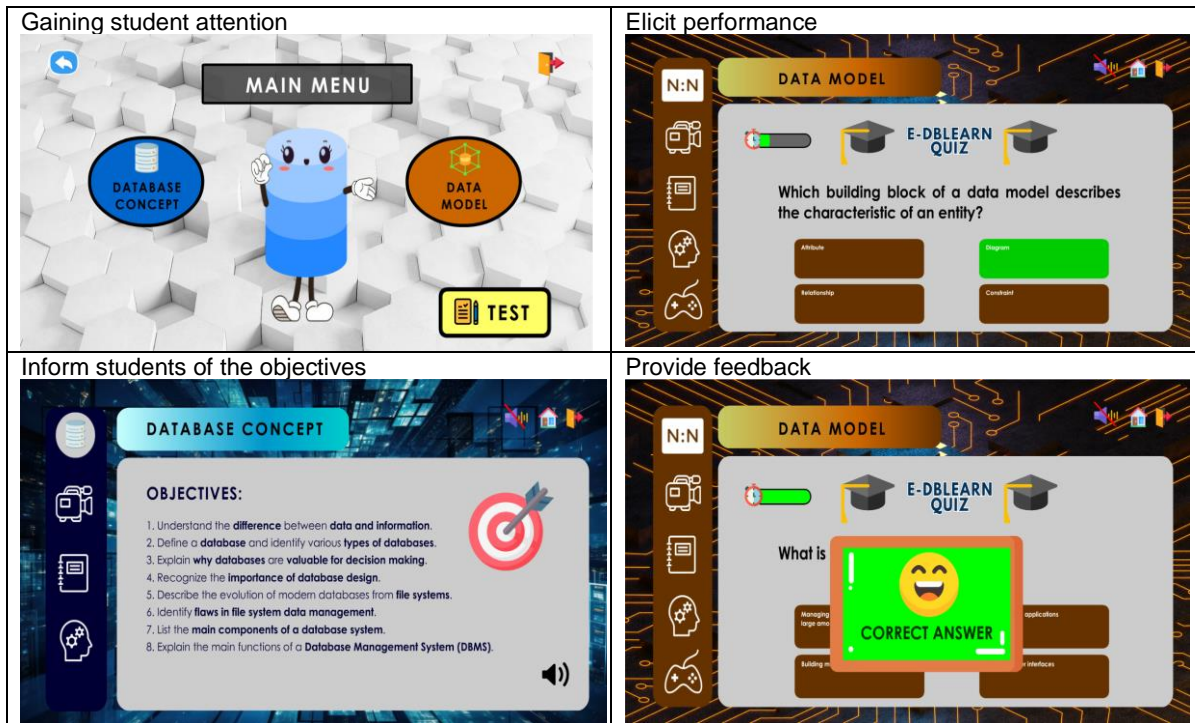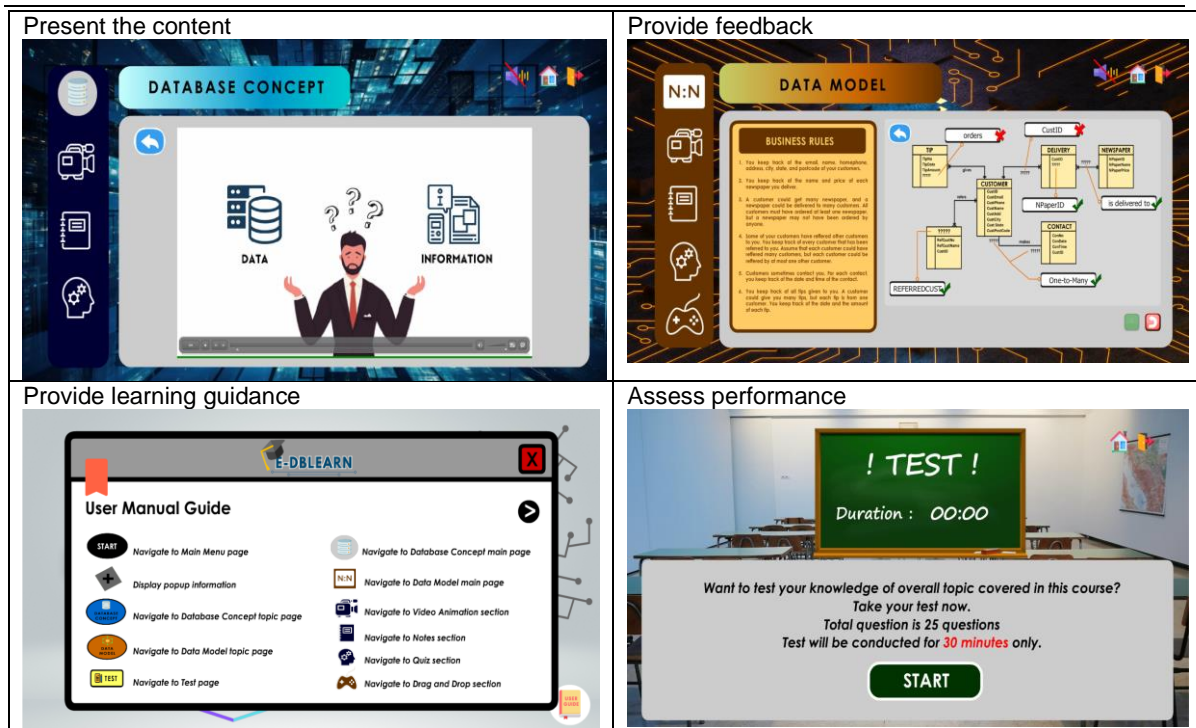
Figure 2. Interface snapshots of E-DBLEARN based on Gagne's Model

In the Implementation phase, the E-DBLEARN application underwent testing according to a comprehensive test plan to ensure its functionality and resolution of technical issues. After successful testing, the application was published as a .exe file and deployed to the target audience, marking its transition from development to ready to use application for further evaluation.

The Evaluation phase assessed the E-DBLEARN application's effectiveness and usability through expert evaluations, with findings informing future updates to maintain its educational value, and the results are presented in Section 4.

4. Results and Discussion

Functionality testing was conducted by both developers and testers. This test included various components of the application: the Start Page, Main Menu Page, Database Concept and Data Model Topics Main Menu Page, Video Section, Notes Section, Quiz Section, Games Section, Test Page, Exit Page, and all associated features. The results of this testing validated that all functionalities in E-DBLEARN were successful and operating as intended.

The expert evaluation of the E-Learning Application for Introduction to Database (E-DBLEARN) utilized a set of questionnaires based on eight criteria of usability evaluation as previously studied by [18]. These criteria were integrated with seven selected characteristics from Gagne's Nine Events of Instruction. The evaluation employed a Likert scale to measure usability, allowing for a nuanced assessment of each criterion, with options spanning from 5 (Strongly Agree) to 1 (Strongly Disagree), to gauge the experts' opinions and perceptions on the e-learning application's usability. The evaluators for the E-DBLEARN expert evaluation are three senior lecturers with over five years of experience in their respective Information Technology (IT) fields: database subjects, multimedia, and human-computer interaction.

According to the evaluation results, presented in Figure 3 to 10, all three experts rated every criterion with a score of 5 (Outstanding). This consistently high rating reflects that E-DBLEARN has been verified in the usability aspects of the application, affirming its effectiveness in delivering an optimal user experience. For the first usability criteria is content. The evaluation of the content aimed to assess the quality and relevance of the educational material in E-DBLEARN, ensuring it supports learning objectives and meets the target audience's needs. Experts' feedback verified that the content is well-organized, accurate, and effectively aligns with the educational goals.

SECTION A: CONTENT		
No.	Description	Expert Rate
1.	Appropriateness of language and terms used	Expert 1: 5 Expert 2: 5 Expert 3: 5
2.	Quality and relevance of learning and supporting materials	Expert 1: 5 Expert 2: 5 Expert 3: 5
3.	Accuracy and completeness of information provided	Expert 1: 5 Expert 2: 5 Expert 3: 5
Comment		Suggestion
Expert 3: Everything okay		-

Figure 3. Result Expert Evaluation for Content

The second section is learning and support. This evaluation focuses on the effectiveness of the application's learning support features, including guidance, help resources, and feedback mechanisms. The experts found E-DBLEARN to excel in this area, highlighting its clear guidance and helpful resources that significantly enhance the user's learning experience.

SECTION B: LEARNING AND SUPPORT		
No.	Description	Expert Rate
1.	Effectiveness of delivering interactive contents (video animations, notes, quizzes, drag-and-drop games, test)	Expert 1: 4 Expert 2: 5 Expert 3: 5
2.	Quality of academic discussions and communication tools	Expert 1: 5 Expert 2: 5 Expert 3: 5
3.	Adequacy of learning assessment methods	Expert 1: 5 Expert 2: 4 Expert 3: 5
Comment		Suggestion
Expert 2: Comply to all multimedia elements		Expert 3: Need to add instruction more as user guidance (quiz)

Figure 4. Result Expert Evaluation for Learning and Support

The third section is visual design. The purpose of assessing visual design was to evaluate the application's layout, design consistency, and overall aesthetic appeal to ensure a positive user experience. The experts gave high ratings for visual design, praising E-DBLEARN for its clear, consistent, and engaging visual elements.

SECTION C: VISUAL DESIGN		
No.	Description	Expert Rate
1.	Clarity and comprehensibility of the interface	Expert 1: 5 Expert 2: 5 Expert 3: 5
2.	Effectiveness of layout, colour scheme, typography and multimedia elements (images, videos, audios and animations)	Expert 1: 5 Expert 2: 5 Expert 3: 5
3.	Overall aesthetic appeal and user engagement	Expert 1: 5 Expert 2: 5 Expert 3: 5
Comment		Suggestion
Expert 1: Add few more animation elements in courseware		Expert 2: Minor improvement to the layout design

Figure 5. Result Expert Evaluation for Visual Design

The fourth section is navigation, which aimed to determine how easily users can move through the application and access its features. Experts rated this section as excellent, noting that E-DBLEARN provides an intuitive and user-friendly navigation system, allowing for efficient access to various sections.

SECTION D: NAVIGATION		
No.	Description	Expert Rate
1.	Ease of browsing and locating information	Expert 1: 5 Expert 2: 5 Expert 3: 5
2.	Intuitiveness of the navigation structure and flow	Expert 1: 5 Expert 2: 5 Expert 3: 5
3.	Efficiency of accessing and utilizing features (quizzes, games, summative tests)	Expert 1: 5 Expert 2: 4 Expert 3: 5
Comment		**Suggestion**
-		Expert 2: Add label in execute answer for quiz

Figure 6. Result Expert Evaluation for Navigation

The fifth section is accessibility. It focuses on how well the application supports users with varying abilities, including those with disabilities. The experts rated E-DBLEARN highly for accessibility, indicating that it incorporates effective features for a broad range of users, including compatibility with assistive technologies.

SECTION E: ACCESSIBILITY		
No.	Description	Expert Rate
1.	Ease of accessing the application across various device	Expert 1: 5 Expert 2: 5 Expert 3: 5
2.	Consistency of the user experience across different platforms	Expert 1: 5 Expert 2: 5 Expert 3: 5
3.	Compliance with accessibility standards	Expert 1: 5 Expert 2: 5 Expert 3: 5
Comment		**Suggestion**
-		-

Figure 7. Result Expert Evaluation for Accessibility

Next section is interactivity assessment sought to review the effectiveness of engaging features such as quizzes and simulations in the application. Experts found E-DBLEARN's interactive elements outstanding, noting that they effectively engage users and reinforce learning.

SECTION F: INTERACTIVITY		
No.	Description	Expert Rate
1.	Effectiveness of interactive tools within the e-learning application	Expert 1: 5 Expert 2: 5 Expert 3: 5
2.	Quality and engagement level of interactive learning activities (quizzes, drag-and-drop games, summative tests)	Expert 1: 5 Expert 2: 5 Expert 3: 5
3.	Responsiveness and reliability of interactive elements	Expert 1: 5 Expert 2: 5 Expert 3: 5
Comment		**Suggestion**
-		-

Figure 8. Result Expert Evaluation for Interactivity

For self-assessment and learnability, this evaluation aimed to assess the application's support for self-assessment and ease of learning, including tools for tracking progress and mastering the system. Experts rated this section highly, noting that E-DBLEARN provides effective self-assessment tools and is easy for users to learn.

SECTION G: SELF-ASSESSMENT AND LEARNABILITY		
No.	**Description**	**Expert Rate**
1.	Quality of self-assessment tools	Expert 1: 5 Expert 2: 5 Expert 3: 5
2.	Ease of learning how to use the e-learning application	Expert 1: 5 Expert 2: 5 Expert 3: 5
3.	Support for autonomous learning and user guidance	Expert 1: 5 Expert 2: 5 Expert 3: 5
Comment		**Suggestion**
-		-

Figure 9. Result Expert Evaluation for Self-Assessment and Learnability

Lastly, the motivation to learn evaluation focuses on how well the application engages users and fosters interest in learning. Experts gave excellent ratings, highlighting that E-DBLEARN effectively motivates users through its design and content, encouraging continued use and learning.

SECTION H: MOTIVATION TO LEARN		
No.	**Description**	**Expert Rate**
1.	Ability to engage and motivate learners through interactive content and multimedia elements	Expert 1: 5 Expert 2: 5 Expert 3: 5
2.	Support for setting and achieving learning goals	Expert 1: 5 Expert 2: 5 Expert 3: 5
3.	Overall enjoyment and satisfaction with the learning experience	Expert 1: 5 Expert 2: 5 Expert 3: 5
Comment		**Suggestion**
Expert 2: Well developed e-learning		

Figure 10. Result Expert Evaluation for Motivation to Learn

In conclusion, the functionality and usability testing of the E-DBLEARN demonstrated its effectiveness in delivering a high-quality learning experience. The comprehensive testing process, conducted by both developers and testers, ensured that all features operated as intended. Additionally, the expert evaluation, based on usability criteria and Gagne's instructional principles, confirmed that E-DBLEARN excels in critical areas such as content quality, learning support, visual design, navigation, accessibility, interactivity, self-assessment, and user motivation. With consistently high ratings from experienced lecturers, E-DBLEARN has been validated as an effective tool for facilitating database education, providing an engaging and accessible learning environment.

5. Conclusion

The E-DBLEARN e-learning application is developed as an assisting tool for teaching and learning of Database Environment and Data Model topic in any database courses. The problem that triggers development stems from the challenges faced by the learners in understanding the complex and interconnected terminology in database environments and data modeling, thus leading to difficulties in applying key concepts effectively. In addition, a preliminary study has revealed that inadequate classroom facilities, varying learning styles, and the use of external resources that are not tailored to the syllabus have negatively impacted the understanding of theory-based topics. The significance of the E-DBLEARN is that it provides interactive multimedia content based on the Gagne Instructional Design Model for the database course, which can be an effective additional learning material for students. This can contribute to the enhancement of both teaching and learning by helping students better understand and apply complex database concepts. Evaluation results from the expert feedback highlighted areas for potential improvements, including the addition of animations, enhanced user guidance and minor layout adjustments. The experts' suggestions offer valuable insights for further refinement, aiming to boost user engagement and interface intuitiveness. In summary, the expert evaluation confirms that E-DBLEARN meets its designated functions.

Conflict of Interest
The authors declare there is no conflict of interest in the subject matter or materials discussed in this manuscript.

References

[1] D. R. Garrison and N. D. Vaughan, Blended Learning in Higher Education: Framework, Principles, and Guidelines. John Wiley & Sons, 2020.

[2] R. E. Mayer, Multimedia Learning, 3rd ed. Cambridge, U.K.: Cambridge Univ. Press, 2021.

[3] R. C. Clark and R. E. Mayer, E-Learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning, 4th ed. Hoboken, NJ, USA: Wiley, 2016.

[4] A. Silberschatz, H. F. Korth, and S. Sudarshan, Database System Concepts, 7th ed. New York, NY, USA: McGraw-Hill, 2019.

[5] C. Coronel and S. Morris, Database Systems: Design, Implementation, & Management, 14th ed. Boston, MA, USA: Cengage Learning, 2022.

[6] D. Al-Fraihat, M. Joy, J. Sinclair, and M. Alsheraideh, "Evaluating E-learning systems success: An empirical study," Computers in Human Behavior, vol. 102, pp. 67-86, 2020.

[7] B. Means, M. Bakia, and R. Murphy, Evaluation of Evidence-Based Practices in Online Learning: A Meta-Analysis and Review of Online Learning Studies, U.S. Department of Education, 2014.

[8] E. F. Barkley, "Interactive Learning in Asynchronous Online Classes," in Collaborative Learning Techniques: A Handbook for College Faculty, 2nd ed. San Francisco, CA: Jossey-Bass, 2020, pp. 230-256

[9] G. Siemens, Learning Analytics: The Emergence of a Discipline, American Behavioral Scientist, vol. 57, no. 10, pp. 1409-1420, 2013.

[10] M. Laufer et al., "Digital higher education: a divider or bridge builder? Leadership perspectives on edtech in a COVID-19 reality," International Journal of Educational Technology in Higher Education, vol. 18, no. 1, 2021.

[11] A. A. Hussain, M. R. Iqbal, and M. H. Khan, "The Impact of Multimedia in E-Learning: A Comprehensive Review," *Educational Technology Research and Development*, vol. 71, no. 1, pp. 115-138, 2023.

[12] R. C. Hwang, J. Y. Chang, and C. H. Lin, "Designing Interactive Multimedia Learning Environments: A Review of Recent Advances," *Journal of Educational Technology & Society*, vol. 26, no. 2, pp. 1-17, 2023.

[13] J. L. Moreno and R. E. Mayer, "Interactive Multimedia Learning Environments: A Meta-Analysis," *Journal of Educational Psychology*, vol. 112, no. 4, pp. 675-689, 2020.

[14] P. Rob and C. Coronel, *Database Systems: Design, Implementation, & Management*, 13th ed. Cengage Learning, 2022.

[15] S. B. Navathe and R. Elmasri, *Fundamentals of Database Systems*, 7th ed. Pearson, 2017.

[16] J. Han, E. Haihong, G. Le, and J. Du, "Survey on NoSQL Database," *Journal of Software*, vol. 8, no. 3, pp. 1-8, Mar. 2021.

[17] A. Razali, "Using Gagné's model of instructional design to develop an engaging learning experience," International Journal of Educational Technology in Higher Education, vol. 21, no. 1, pp. 34-45, 2024.

[18] P. Zaharias and A. Poylymenakou, "Developing a usability evaluation method for e-learning applications: Beyond functional usability," *Int. J. Hum.-Comput. Interact.*, vol. 25, no. 1, pp. 75-98, 2009. DOI: 10.1080/10447310802546716.

Biography of all authors

Picture	Biography	Authorship contribution
	Fatin Najwa Kamal is a final year student of Bachelor Information Systems (Hons.) Business Computing	Data collection, data analysis and drafting article.
	Nor Azila Awang Abu Bakar is an academic staff in UiTM Terengganu, Kuala Terengganu branch.	Designing the research work and reviewing the article.
	Siti Nurul Hayatie Ishak is an academic staff in UiTM Terengganu, Kuala Terengganu branch.	References and formatting.
	Dr. Norlina Mohd Sabri is an academic staff in UiTM Terengganu, Kuala Terengganu branch.	Content editing.

Application Of Cubic Trigonometric Hermite Interpolation Curve with Shape Parameters In 2 Dimensional Objects

Nursyazni Mohamad Sukri
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Cawangan Terengganu
Kampus Kuala Terengganu, Malaysia
nursyazni@uitm.edu.my

Farah Nabilah Azmi
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Cawangan Terengganu
Kampus Kuala Terengganu, Malaysia
farahnabilahofficial@gmail.com

Noor Khairiah Razali
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Cawangan Terengganu
Kampus Kuala Terengganu, Malaysia
noorkhairiah@uitm.edu.my

Siti Musliha Binti Nor-Al-Din
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Cawangan Terengganu
Kampus Kuala Terengganu, Malaysia
Sitim907@uitm.edu.my

Article Info

ABSTRACT

Achieving accurate mathematical representations of an object's geometry for computer-aided design and manufacturing is a complex challenge. Designing such objects requires a method that offers the necessary flexibility and smoothness in curves, which traditional approaches often lack.. The primary aim of this study is to explore and implement cubic trigonometric Hermite interpolation as a method for constructing two-dimensional objects. Additionally, this study seeks to examine the properties of this interpolation method and determine the best value of the shape parameter that produces the smoothest curves. Cubic trigonometric Hermite interpolation is used for constructing two-dimensional objects, with the process carried out using Mathematica. Different values of the shape parameter are employed to achieve the desired curve characteristics in the geometric models. The correct images of the two-dimensional objects are also located, and the suitable shape parameters can be obtained after studying the behavior of the curve using free parameters. Two dimensional objects can be formed using cubic trigonometric Hermite interpolation without making any easy blunders or errors. This project has significant potential to enhance geometric modeling by determining the optimal shape parameters that improve the precision and efficiency of creating two-dimensional geometric models. The results can greatly benefit technical drawing, computer-aided design, and manufacturing by providing a more flexible and accurate method for representing the geometry of objects.

Corresponding Author:
Nursyazni Binti Mohamad Sukri
College of Computing, Informatics, and Mathematics, Universiti Teknologi MARA Terengganu, Kampus Kuala Terengganu, Terengganu, Malaysia
Email: nursyazni@uitm.edu.my

1. Introduction

Curved objects are designed, computed, and digitally represented in Computer-Aided Geometric Design (CAGD) [1]. Thus, it's unsurprising that CAGD has historically been closely linked to traditional mathematical fields. These include approximation theory (emphasizing polynomial and

piecewise polynomial functions), differential geometry (focusing on parametric surfaces), algebraic geometry (concerning algebraic surfaces), functional analysis and differential equations (for surface design through minimization of functionals), and numerical analysis. Vlachkova in 2020, identifies the interpolation of scattered data as a critical issue in both approximation theory and CAGD, noting its application across various industries such as architecture, computer graphics, and the design of ships and airplanes [2].

Numerous multivariate interpolation algorithms have been introduced. A vital element in modeling interpolation curves is the cubic Hermite interpolation curve, which, however, suffers from three major drawbacks: poor continuity, difficult shape adjustment, and an inadequate capability to accurately represent several common engineering curves [3]. To overcome these limitations, the cubic trigonometric Hermite interpolation curve has been developed. This curve accurately represents the elliptical arc, circular arc, quadratic parabolic arc, cubic parabolic arc, and astroid arc that frequently occur in engineering designs. It also achieves C2 continuity and offers both local and global adjustability [4].

Therefore, the primary goal of this project is to explore the construction of two-dimensional objects using cubic trigonometric Hermite interpolation. Curves are generated using the same control point across different methods. The resulting curves are then examined, and their behaviors are analyzed. Furthermore, this project introduces the concept of variable parameters that can control the characteristics of the cubic trigonometric Hermite interpolation, including adjustments in the coordinates and the tangents.

2. Literature Review

Cubic Hermite parametric curves and surfaces have been widely applied in Computer Aided Graphic Design (CAGD). Several authors have explored the Hermite trigonometric interpolation problem. Han in 2015, presented explicit piecewise trigonometric Hermite interpolation methods based on the symmetric, nonnegative, and normalized basis of trigonometric polynomials. These trigonometric Hermite interpolants are local and easy to compute [5],[6].

The bivariate rational Hermite interpolation method was developed in 2008 to create a space surface using the function values and the function's first-order partial derivatives as the interpolation data. Additionally, an explicit and simple mathematical expression for the interpolation function with a parameter was provided [7].

A new class of cubic trigonometric Hermite parametric curves and surfaces with shape parameters was introduced in 2012, similar to regular cubic Hermite parametric curves and surfaces. These new surfaces and curves inherit the properties of standard cubic Hermite surfaces and curves in polynomial space while also offering additional beneficial modeling qualities [8].

Furthermore, Li and Liu [4] proposed optimization solutions based on internal energy minimization for the cubic trigonometric Hermite interpolation curve, aiming to optimize the shapes of planar and spatial curves. Several modeling examples demonstrated the success of these methods, showing that cubic trigonometric Hermite interpolation curves are more practical than cubic Hermite interpolation curves [4].

Veldin et al. [9] presented a discrete nonlinear Kirchhoff-Love four-node shell finite element based on the bi-linear Coons surface patch. This patch crosses the region between the edge curves of cubic Hermites and their edge curves, producing cubic Hermite edge curves by decreasing the bending curvature of a spatial curve that connects two neighboring nodes of the element [9]. In 2023, Žagar examined the problem of interpolating two locations, two tangent directions, curvatures, and the arc length sampled from a circular arc (circular arc data). Planar Pythagorean-hodograph (PH) curves with a pitch of seven degrees were preferred due to their ample free parameters and ease of interpolating the arc length [10].

Rabbath and Corriveau [11] compared the potential curve fitting techniques of Piecewise Cubic Hermite Interpolating Polynomial (PCHIP), cubic splines, and piecewise linear functions. They introduced the procedures required to generate such piecewise polynomial functions using accessible tools to approximate the aerodynamics of a generic small arms projectile. Their contribution included evaluating the projectile aerodynamics approximation using PCHIP and offering an initial assessment of how the polynomial functions affect flight trajectory forecasts derived from 6-degree-of-freedom simulations of a standard projectile [11].

Han and Yang [12] presented a two-step modeling method for developing a smoothing curve or surface representation model for interpolating interval data. The first step involves developing a cubic Hermite spline model to create a smoothing piecewise parametric curve of continuity that

reveals the implicit relationship between variables. In the second stage, sample points are selected from the best parametric spline curve from the first step as interpolation points to build explicit expressions and demonstrate the direct relationship between variables [12].

Albrecht et al. [13] addressed the G2 and C1 Hermite interpolation problem using planar quintic Pythagorean Hodograph B-spline curves with a single free internal knot serving as a shape parameter. This involves interpolating prescribed boundary locations, first derivatives, and curvatures at these points. They provided prerequisites for the data that guarantee solutions and demonstrated how the internal knot affects the values of the absolute rotation index or bending energy, as well as the shape of the resulting interpolant [13],[14].

The Hermite interpolation problem is solved by Arnal et al. [15] by passing the unique rationally parameterized curve of constant width through certain user-controlled points and tangents, provided an allowable denominator and a width value. The outcome provides an easy explanation for such a parameterization and is constructive. Thus, by selecting a set of points with their related tangents, these curves can be designed in a dynamic and participatory manner.

3. Methodology

3.1 Cubic Trigonometric Hermite Interpolation curve

The Cubic Trigonometric Hermite is used to form the two dimensional objects. All the calculations are calculated using Mathematica software. Relying on the research made by Li and Liu in 2022, the cubic trigonometric Hermite interpolation curve is afterwards obtained naturally. The equation (1) shows the basis function of cubic trigonometric Hermite interpolation curve [4]. Figure 1 shows the basis function for cubic trigonometric interpolation.

$$F_{i,0}(t) = (1-\alpha_i) + 3(\alpha_i - 1)S^2 + 2(1-\alpha_i)S^3 + \alpha_i C^3$$

$$F_{i,1}(t) = 2(\alpha_{i+1} - 1) + 3(1-\alpha_{i+1})S^2 + \alpha_{i+1}S^3 + 2(1-\alpha_{i+1})C^3$$

$$G_{i,0}(t) = -\beta_i + S + 3(\beta_i - 2)S^2 + (1-2\beta_i)S^3 + \beta_i C^3$$

$$G_{i,1}(t) = \frac{1}{3}\left(2(3\beta_{i+1} - 5) - 3C + 9(2-\beta_{i+1})S^2 + 3(\beta_{i+1} - 8)S^3 + (13-6\beta_{i+1})C^3\right)$$

(1)

where $S = \sin(t)$, $C = \cos(t)$, $t \in \left[0, \frac{\pi}{2}\right]$, α_i, α_{i+1}, β_i, and β_{i+1} are shape parameters.
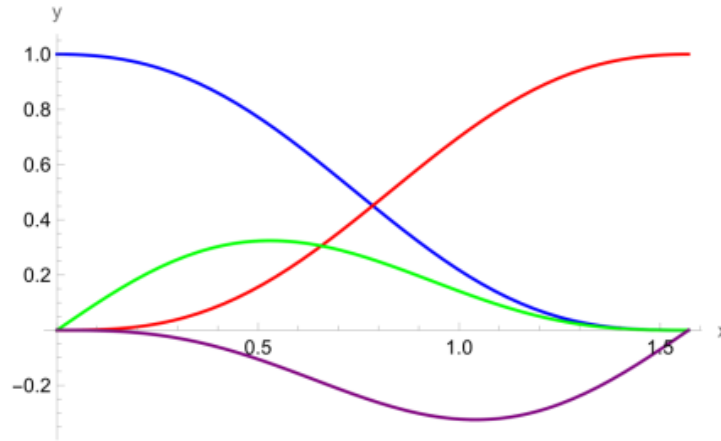

Figure 1. The basis function graph for cubic trigonometric Hermite interpolation

As shown in Figure 1, red curve is formed using $F_{i,0}(t)$, blue curve is formed using $F_{i,1}(t)$. $G_{i,0}(t)$ is used to form green curve while $G_{i,1}(t)$ is used to formed purple curve. Then, all the basis functions are used to form the cubic trigonometric Hermite interpolation curve, also known as the CTHI curve.

3.2 Curve behavior of Cubic Trigonometric Hermite Interpolation

The cubic trigonometric Hermite interpolation curve is defined by equation (2) given a set of points p_j and the corresponding tangent vectors m_j;

$$TH_i = F_{i,0}(t)\,p_i + F_{i,1}(t)p_{i+1} + G_{i,0}(t)\,m_i + G_{i,1}(t)\,m_{i+1} \tag{2}$$

where $i = 0,1,.....,n-1$, $0 \le t \le 1$,

In this paper, point, $p_0 = (0,0)$ and $p_1 = (1,0)$, while the tangent vector $m_0 = (0,1)$ and $m_1 = (0,1)$ are used to form the curve. Position of the point p_j and the tangent vector m_j are maintained while the four shape parameters α_i, α_{i+1}, β_i, and β_{i+1} are changing in order to change the shape of curve.

Figure 2 shows the difference shape of curves obtained if difference values of shape parameters are used. The cubic trigonometric Hermite interpolation curve can be modified by changing the shape parameter value. Curve (a) and curve (b) used the same value of control point which are $P0 = P2 = (0,0)$ and $P1 = P3 = (-1,0)$ while the tangent points also are same for all the curves at $T0 = T2 = (1,0)$ and $T1 = T3 = (-1,0)$. Shape parameter that used to form curve (a) is $\alpha_i = -0.5$, $\alpha_{i+1} = 5$, $\beta_i = 1$, $\beta_{i+1} = 0.5$ and curve (b) is $\alpha_i = 0.5$, $\alpha_{i+1} = 0$, $\beta_i = -5$, $\beta_{i+1} = -1$, while shape parameter for curve (c) is $\alpha_i = 0.5$, $\alpha_{i+1} = 1$, $\beta_i = 0$, $\beta_{i+1} = -0.5$ and the shape parameter for curve (d) is $\alpha_i = -1$, $\alpha_{i+1} = -0.5$, $\beta_i = 0$, $\beta_{i+1} = 0.5$.

(a)

(b)

(c)

(d)
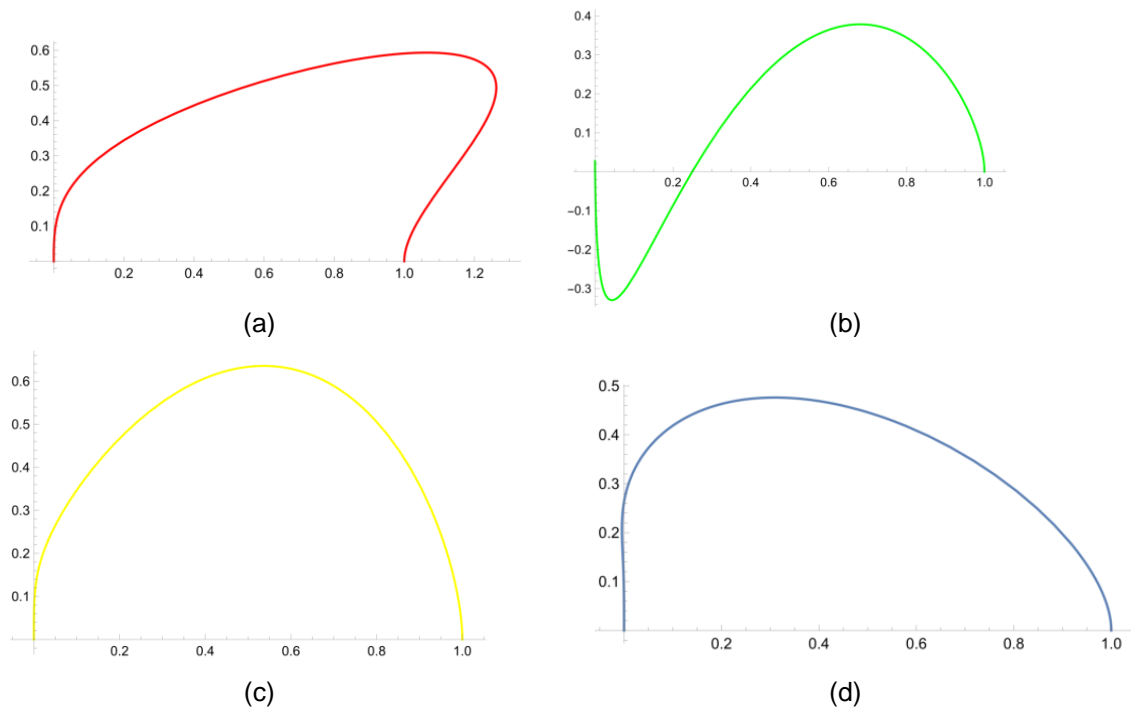
Figure 2. Variety shape of curve obtained by using difference value of shape parameter

4. Results and Discussion

4.1 Application of Cubic Trigonometric Hermite Interpolation curve in 2-dimensional

The Cubic Trigonometric Hermite Interpolation method is used to design the 2-dimensional shape which are crescent and circle. Both curve must be on the same quadrant in order to design the crescent. While to design the circle, curve 1 must be on the quadrant that opposite to the curve 2. Then, the control point, the tangent point and shape parameter are need to modify in order to produce desired shape.
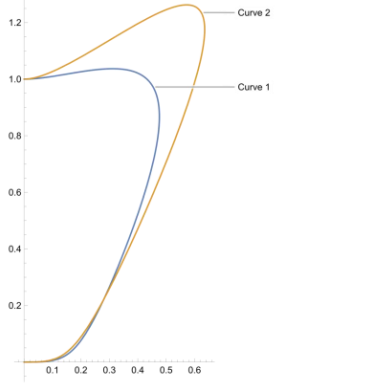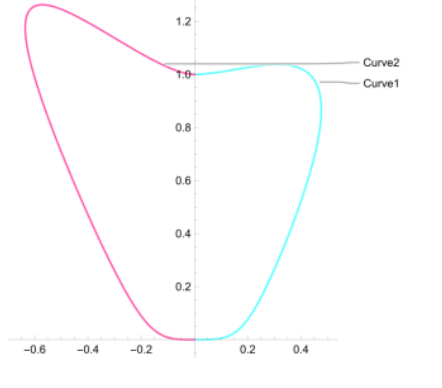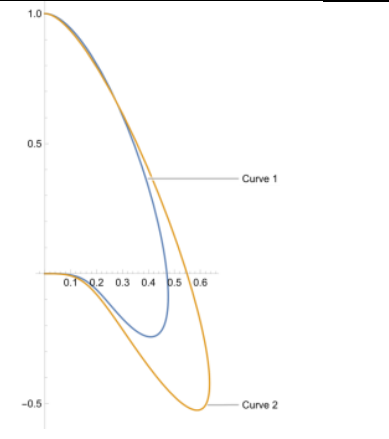
Figure 3 (a) shows that the curve 2 is on the first quadrant when the control points $P2 = (0,0)$ and $P3 = (-1,0)$ while the tangent point $T2 = (1,0)$ and $T3 = (-1,0)$ are used. Figure (b) shows that the curve 2 appeared on the second quadrant when the control points $P2 = (0,0)$ and $P3 = (-1,0)$ while the tangent point $T2 = (0,1)$ and $T3 = (0,-1)$ are used.

Table 1 shows the curves produced when each of the shape parameters is changing respectively from the value of curve 1: $\alpha_i = -1, \alpha_{i+1} = -0.5, \beta_i = 0, \beta_{i+1} = 0.5$ and curve 2: $\alpha_1 = 0.5, \alpha_2 = 1, \beta_1 = 0, \beta_2 = -0.5$.
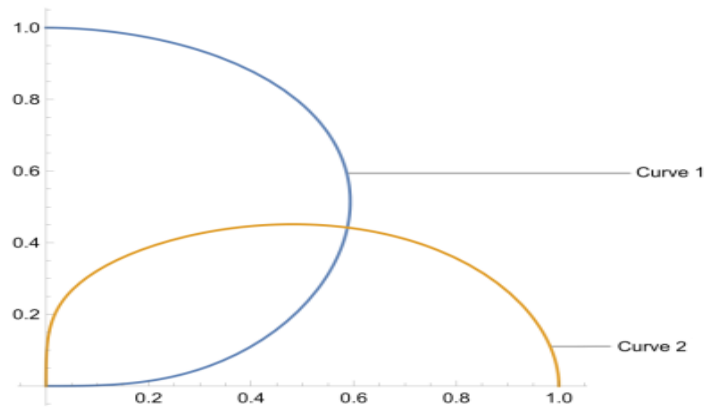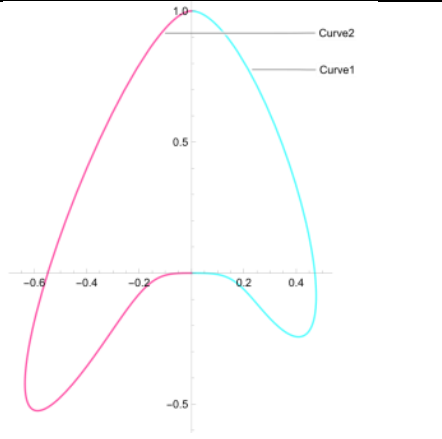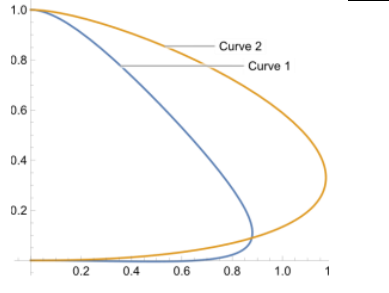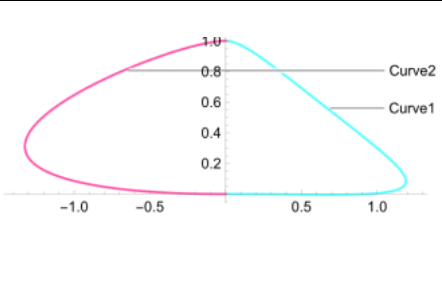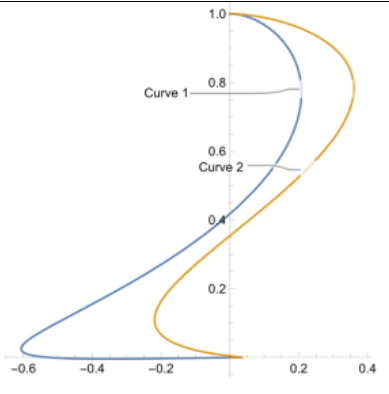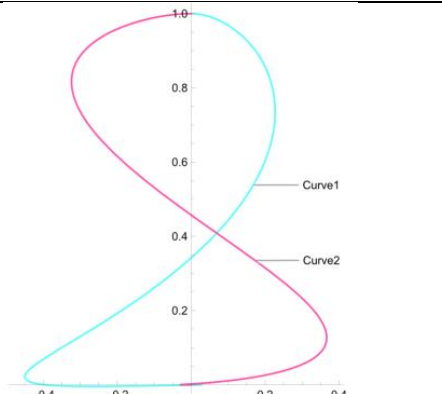

(a)
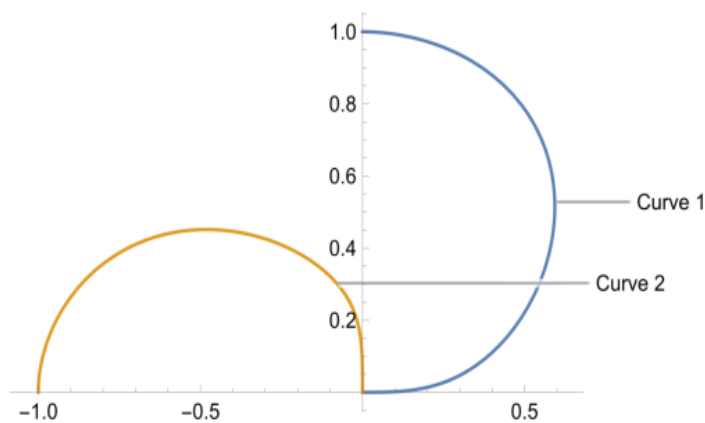
(b)

Figure 3. Position of the curve 2 using different control point and tangent point.

Table 1. The curve obtained when each of the shape parameter is changing respectively

		The value of shape parameter α_{i+1} for curve 1 and α_2 for curve 2 are changed to positive number while other parameters have remained same, the curve 1 and curve 2 will curve upward.
		The value of shape parameter α_{i+1} for curve 1 and α_2 for curve 2 are changed to negative number while other parameters have remained same, the curve 1 and curve 2 will curve downward.
		The curves formed when the value of β_i and β_1 are change to positive number.
		The curves formed when the value of β_i and β_1 are change to negative number.

		The curves formed when the value of β_{i+1} and β_2 are change to positive number.
		The curves formed when the value of β_{i+1} and β_2 are change to negative number.
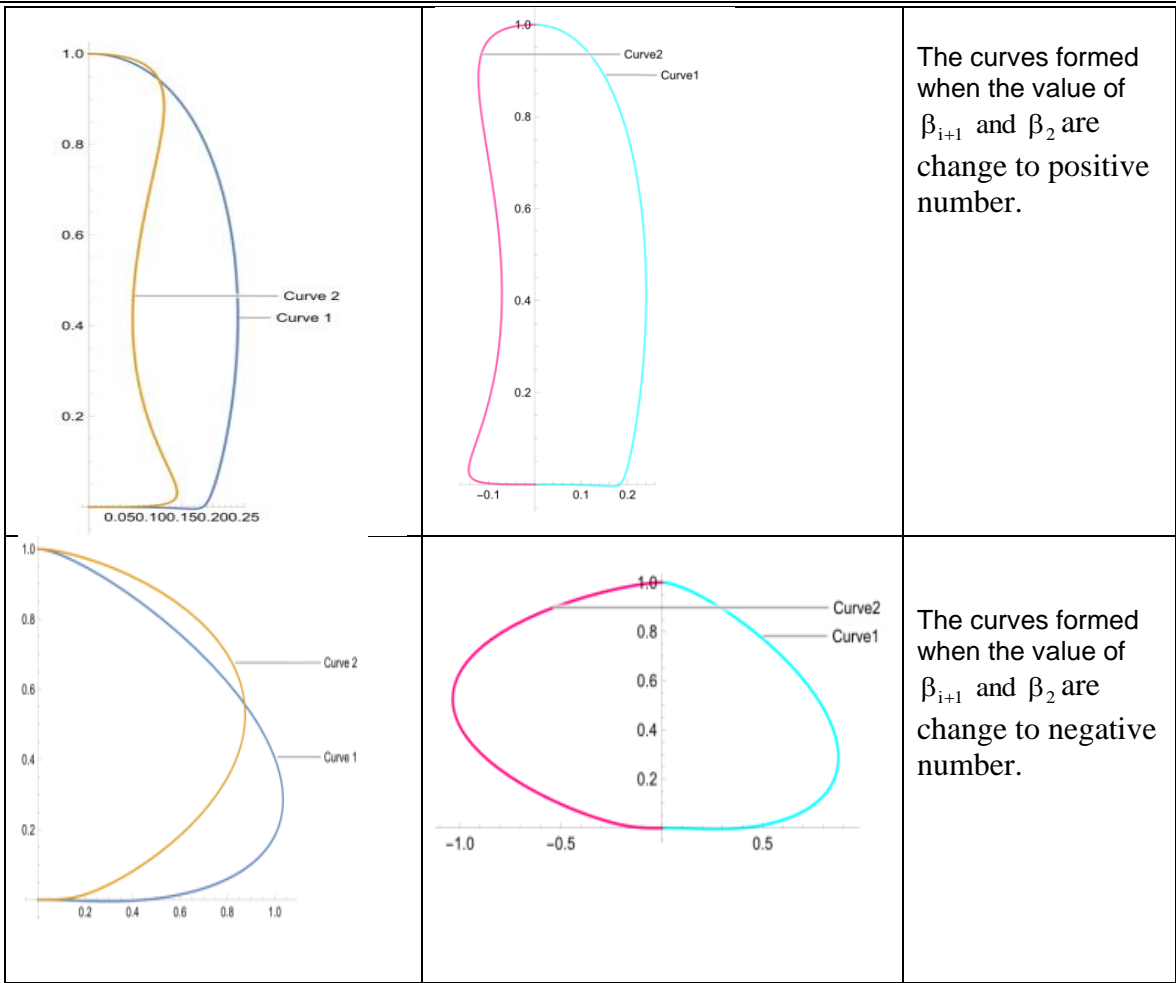
From table 1, the best shape parameters used has been determined to form the smooth 2-dimensional crescent and circle outline. Figure 4 and Figure 5 shows the crescent and circle outline are design using cubic trigonometric Hermite interpolation. The curve has been constructed based on the control point, and the curve has been drawn according to the shape parameters where all the parameters' values are positive in range 0.5 until 2.
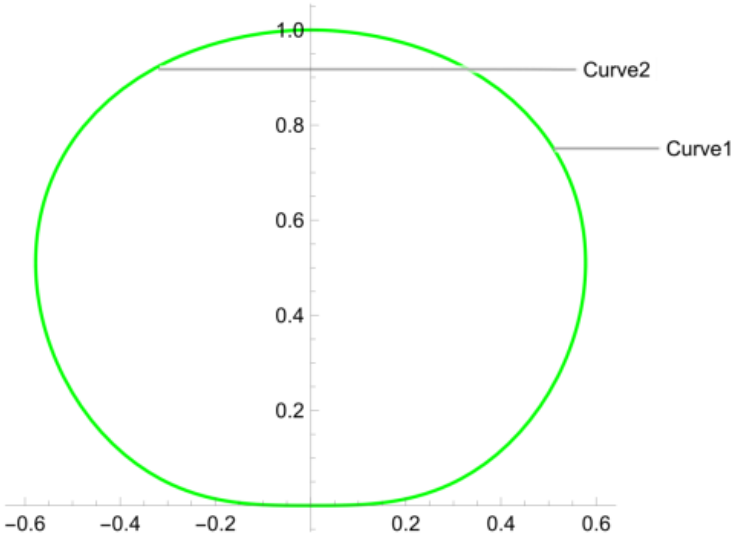

Figure 4. The circle outline formed by using Cubic Trigonometric Hermite Interpolation curves
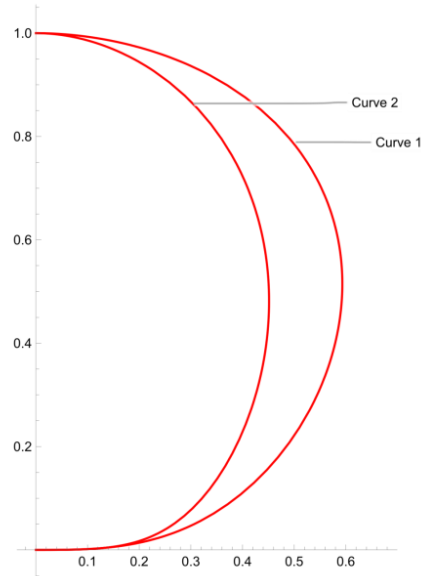
43

Figure 5. The crescent outline formed by using Cubic Trigonometric Hermite Interpolation curves

For crescent outline, the control points for curve 1 and curve 2 are same at $P0 = P2 = (0,0)$ and $P1 = P3 = (-1,0)$ while the tangent points for curve 1 and curve 2 also are same at $T0 = T2 = (1,0)$ and $T1 = T3 = (-1,0)$. The best shape parameters for each curve have been chosen to generate a two-dimensional crescent shape are shown as in Table 2 below.

Table 2. Shape parameter values for Crescent Outline shape

Curve	α_i	α_{i+1}	β_i	β_{i+1}
Curve 1	2	1.5	1	0.5
Curve 2	2	1	0.5	1

While for the circle outline, the control points for curve 1 and curve 2 are same at $P0 = P2 = (0,0)$ and $P1 = P3 = (-1,0)$ while the tangent points for curve 1 $T0 = (1,0)$ and $T1 = (-1,0)$ and curve 2 is at $T2 = (-1,0)$ and $T3 = (1,0)$. The best shape parameters for each curve have been chosen to generate a two-dimensional crescent shape are shown as in Table 3.

Table 3. Shape parameter values for Circle Outline shape

Curve	α_i	α_{i+1}	β_i	β_{i+1}
Curve 1	2	1.5	1	0.6
Curve 2	2	1.5	1	0.6

5. Conclusion

The findings of this study enable us to reach the conclusion that every one of the research goals has been well achieved. The correct images of the two dimensional objects are also located and the suitable shape parameters can be obtained after studying the behavior of the curve using free parameters.

The behavior of the curve can be changed by adjusting the curve's shape parameters. Although a smooth curve is generated, if some of the shape parameters are modified, a sharp curve will be generated. A relatively small difference between the parameters of two curves may be discernible. The shape that is assigned to the curves constructed using the cubic trigonometric Hermite interpolation function depends on the shape parameters of the shape that is created with

the control points and tangent vectors. When it comes to the creation of two-dimensional shapes, the cubic trigonometric Hermite interpolation function results in smooth curves. It is possible to decide the suitable shape parameters of the cubic trigonometric Hermite interpolation function. Additionally, using the appropriate shape parameters is the only way to produce tidy and respectable results. Because each shape has a unique combination of curve features, using the software is quite helpful when undertaking research that involves the development of two-dimensional shapes.

Conflict of Interest
The authors declare no conflict of interest in the subject matter or materials discussed in this manuscript.

References

[1]   S. Kazem and A. Hatam, "Journal of Computational and Applied Scattered data interpolation: Strictly positive definite radial basis / cardinal functions," Journal of Computational and Applied Mathematics, vol. 394, p. 113580, 2021, doi: 10.1016/j.cam.2021.113580.

[2]   K. Vlachkova, "Interpolation of scattered data in R3 using minimum Lp - norm," Journal of Mathematical Analysis and Applications, vol. 485, no. 2, p. 123824, 2020, doi: 10.1016/j.jmaa.2019.123824.

[3]   C. A. Rabbath and D. Corriveau, "A comparison of piecewise cubic hermite interpolating polynomials, cubic splines and piecewise linear functions for the approximation of projectile aerodynamics," in Proceedings - 31st International Symposium on Ballistics, BALLISTICS 2019, 2019, vol. 1, pp. 610–614, doi: 10.12783/ballistics2019/33099.

[4]   J. Li and C. Liu, "Cubic Trigonometric Hermite Interpolation Curve: Construction, Properties, and Shape Optimization," Journal of Function Spaces, vol. 2022, 2022, doi: 10.1155/2022/7525056.

[5]   X. Han, "Piecewise trigonometric Hermite interpolation," Appl. Math. Comput., vol. 268, pp. 616–627, 2015.

[6]   F. Oumella and A. Lamnii, "Curve and Surface Construction Using Hermite Trigonometric Interpolant," Mathematical and Computational Applications, vol. 26, 2021.

[7]   Q. Duan, S. Li, F. Bao, and E. H. Twizell, "Hermite interpolation by piecewise rational surface," Appl. Math. Comput., vol. 198, pp. 59–72, 2008.

[8]   J. Li, "Cubic Trigonometric Hermite Parametric Curves and Surfaces with Shape Parameters," Int. J. Inf. Comput. Sci., vol. 1, p. 8, 2012.

[9]   T. Veldin, B. Brank, and M. Brojan, "Discrete Kirchhoff–Love shell quadrilateral finite element designed from cubic Hermite edge curves and Coons surface patch," Thin-Walled Structures, vol. 168, 2021, doi: 10.1016/j.tws.2021.108268.

[10]   E. Žagar, "Arc length preserving G2 Hermite interpolation of circular arcs," Journal of Computational and Applied Mathematics, vol. 424, 2023, doi: 10.1016/j.cam.2022.115008.

[11]   C. A. Rabbath and D. Corriveau, "A comparison of piecewise cubic hermite interpolating polynomials, cubic splines and piecewise linear functions for the approximation of projectile aerodynamics," in Proceedings - 31st International Symposium on Ballistics, BALLISTICS 2019, 2019, vol. 1, pp. 610–614, doi: 10.12783/ballistics2019/33099.

[12]   X. Han and J. Yang, "A two-step method for interpolating interval data based on cubic hermite polynomial models," Applied Mathematical Modelling, vol. 81, no. 11771453, pp. 356–371, 2020, doi: 10.1016/j.apm.2019.12.013.

[13]   G. Albrecht, C. V. Beccari, and L. Romani, "G2⁄C1 Hermite interpolation by planar PH B-spline curves with shape parameter," Applied Mathematics Letters, vol. 121, pp. 1–8, 2021, doi: 10.1016/j.aml.2021.107452.

[14]   R. Dębski, "Streaming Hermite interpolation using cubic splinelets," Computer Aided Geometric Design, vol. 88, pp. 1–17, 2021, doi: 10.1016/j.cagd.2021.102011.

[15]   A. Arnal, J. V. Beltran, J. Monterde, and D. Rochera, "Geometric Hermite interpolation by rational curves of constant width," Computer and Applied Mathematics, vol. 439, 2024, doi: 10.1016/j.cam.2023.115598.

Biography of all authors

Picture	Biography	Authorship contribution
	Nursyazni Binti Mohamad Sukri is an academic staff of Universiti Teknologi MARA Cawangan Terengganu Kampus Kuala Terengganu in the College of Computing, Informatics and Mathematics. Her research interests are in Computer Aided Graphics Design CAGD and numerical. She can be contacted for further correspondence via email at nursyazni@uitm.edu.my	Introduction, methodology, discussion, and final checking of the article.
	Farah Nabilah Azmi is a student of Universiti Teknologi MARA Terengganu Branch, Kuala Terengganu Campus, Terengganu. She is studying for a bachelor's degree of Science (Hons.) in Mathematical Modelling and Analytics. She can be reached at farahnabilahofficial@gmail.com	Introduction, literature review, data collection, methodology, and analysis.
	Noor Khairiah Razali is an academic staff of Universiti Teknologi MARA Cawangan Terengganu Kampus Kuala Terengganu in the College of Computing, Informatics and Mathematics. Her research interests are in Computer Aided Graphics Design (CAGD), Numerical Method and Ordinary Differential Equation. She can be contacted for further correspondence via email at noorkhairiah@uitm.edu.my	Interpretation and final checking of the article.
	Siti Musliha Binti Nor-al-din is an academic staff of Universiti Teknologi MARA Cawangan Terengganu Kampus Kuala Terengganu in the College of Computing, Informatics and Mathematics. Her research interests are in Applied Optimization. She can be contacted for further correspondence via email at sitim907@uitm.edu.my	literature review and final checking of the article.

Operational Cost Optimization on Solid Waste Collection Using Goal Programming

Diana Sirmayunie Mohd Nasir
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Perlis Branch, Perlis, Malaysia
dianasirmayunie@uitm.edu.my

Salwani Asman
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Perlis Branch, Perlis, Malaysia
2020492712@student.uitm.edu.my

Suzanawati Abu Hasan
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Perlis Branch, Perlis, Malaysia
suzan540@uitm.edu.my

Article Info	ABSTRACT
	Waste collection is the waste management, comprising storage, collection, transportation, and disposal, is the most expensive aspect of waste management, involving resources in strategy, execution, and oversight. Even in Perlis, the smallest state in Malaysia, waste management has become a concern due to population growth and increased industrial activities. This study aims to determine the optimal operational cost by utilizing the Goal Programming approach. Data from E-Idaman Kangar Sdn. Bhd, a private company managing solid waste in Perlis, was utilized. The goal is to achieve an optimization rate by formulating an objective function and utilizing Lingo software version 18.0. The study aims to achieve five objectives regarding operational cost which are minimizing labor cost, minimizing collection cost, minimizing vehicle cost, minimizing consumable cost and minimizing the financial statement of operation for each route. The research results indicate that E-Idaman Kangar's collection expenses are already at an optimal state. It showed that the cost of each vehicle included in the collection activity is optimized. The objective value shows that the total spending can still be reduced by RM3, which is a decrease of 0.01%. In conclusion, this can be seen that the organization accomplishes collection in an ideal manner. The outcomes of this study are poised to contribute towards achieving sustainable Solid Waste Collection (SWC) across the entirety of Perlis state, thereby upholding a clean and healthy environment for its residents. Additionally, through meticulous budgeting and strategic planning prior to each monthly collection, companies can effectively manage their finances and bolster profitability.

Corresponding Author:
Diana Sirmayunie Mohd Nasir
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Perlis Branch, Perlis, Malaysia.
dianasirmayunie@uitm.edu.my

1. Introduction

Solid waste is the term used to describe unwanted materials generated by commercial, industrial, and residential operations. It is significant to recognize that not all waste is considered solid waste. Putting aside the huge amount of waste daily, waste is categorized into a few forms, including food waste, marine litter, and microplastic [1]. Indeed, integrated solid waste management is necessary to reduce the waste-related issue's negative effects on both human health and the environment. This type of management includes waste prevention, recycling,

PENERBIT PRESS
UNIVERSITI TEKNOLOGI MARA

composting, collecting, transportation, burning, and disposal. With an increase in the number of Malaysian populations by 32.8 million, a massive amount of solid waste is being produced, estimated at around 38,427 metric tons per day in the year 2021, which is disposed of at a rate of 82.5 percent in landfills [2]. The amount of municipal solid waste (MSW) that will be collected by 2022 will be over 14 million tons, which is enough to fill the Petronas Twin Tower every seven days. The nature of the solid waste produced has changed because of rapid industrialization and urbanization. The rising rates of annual waste generation are brought on by Malaysians' desire for a higher standard of living [3].

The most expensive part of waste management systems is waste collection. Direct and indirect costs both play a role in the overall expenditure on waste collection. All direct costs made in the management of solid waste in an area are included which covers the resources employed in the planning, implementation, and management of waste management (storage, collection, transportation, and disposal). Meanwhile, indirect costs are the outside expenses involved when using the current waste management techniques which include the cost of hazardous waste collection, storage, and disposal techniques causing environmental harm [4]. Thus, proper, and effective waste management is necessary to maintain a good environment, health, and economic stability. In addition to preserving the environment and public health, research on solid waste disposal methods is necessary to keep costs associated with waste collection and cleaning low.

The smallest state in Malaysia, Perlis Indera Kayangan, is situated in Peninsular Malaysia's northern region. As of 2020, there were 254,700 residents. There are significant environmental issues in Perlis because of the growth in population and industrial activity. One of the issues facing the environment is solid waste. The Environmental Department of Perlis reported Perlis produced 7,816 tons of scheduled waste overall in 2019, up from the 2,536 tons reported in 2018 [5]. To address this environmental issue, the department of cleanliness and the environment conducts waste collection activities.

A collection of solid waste in Perlis is conveyed by E-Idaman Sdn Bhd (E-Idaman). The collection is carried out daily or weekly, while the final disposal of waste from the entire Perlis region takes place at the Padang Siding landfill. There are six main functional components to solid waste management. These include waste production, storage, collection, transport, recycling, processing, and disposal of waste. With the help of the goal programming method, it is hoped that the results of this study will help obtain the optimal cost for solid waste collection.

2. Literature Review

A few studies have been conducted on solid waste collection management. Many different methods were used to obtain optimal time and cost for solid waste collection. As an example, researchers in [6] stated that the state's government invests a sizable sum of money each year in solid waste management but lacks practical means of minimizing or reducing it. It can be inferred from the information above that operational costs are high. Meanwhile, studies from [7] show that solid waste collection can be optimized by investigating route optimization using a Geographical Information System (GIS). Five (5) routes in Ipoh and Perak were selected, and the current route can be optimized as the routes reduce and affect the time needed for waste collection. The analysis was based on the needs and constraints of the collection service. Since then, the GIS-ArcView application for route optimization in Ipoh city has shown a reasonable improvement in route length and travel time reduction.

To facilitate long-term planning of waste management activities in the city of Tianjin, the interval joint-probabilistic mixed-integer programming (IJMP) method is created. Researchers in [8] stated that in the IJMP, few joint-probabilistic constraints were introduced into an interval-parameter mixed-integer programming framework. The uncertainties expressed in terms of interval values and random variables can be considered. The created model may help address the solid waste management system's dynamic, interactive, and uncertain characteristics and can effectively link system cost and system failure risk. Another problem under uncertainty is identifying optimal waste-flow-allocation schemes. A Stepwise Interactive Algorithm (SIA) was used by [9] in a case study of waste allocation problems under uncertainty as an advanced to a generalized fuzzy linear programming (GFLP) method. From the results, reasonable solutions resulted in planning waste allocation practices and the GFLP can provide more detailed information. From there, the decision-makers can construct trade-offs for system stability and plausibility. This can also identify desired policies for any solid waste problems under uncertainty.

Besides, a study has been made to optimize the logistic network and transportation system of Integrated Solid Waste Management (ISWM). This is where it is formulated as a tri-echelon ISWM logistic network. The mixed-integer linear programming (MILP) was used by [10] to formulate the ISWM system in the Fleet Size and Mix Vehicle Routing Problems with the Time Windows framework. This study successfully supports a cost-effective ISWM transportation system, and this proposed approach was successfully applied to real-world problems in southern Tehran, Iran. This approach has obtained the optimal economic cost of the system under uncertainty. The results confirmed the method's strength and effectiveness when the estimated and real amounts of the uncertain parameter differed by larger amounts and when the system network experienced unforeseen disturbances.

Meanwhile, many researchers have studied goal programming, which has been successfully applied to a wide range of real-world problems. [11] proposed the method, which is now accepted as a basic mathematical programming method for solving decision-making problems with multiple objectives. To begin with, the lexicographic goal programming technique was used by [12] to solve the multi-objective optimization model. Three objectives have been considered: Cost reduction, final waste disposal to landfill reduction, and environmental impact reduction are all priorities. As decision makers' preferences, the goal of cost minimization was considered more important than minimizing greenhouse gas emissions and the final amount of waste to the landfill in this problem. The GLPK Integer Optimizer was used to run the model and obtain the optimal solution. The final formulated objective function produces results in which all deviation variables are zero, indicating that all goals are perfectly satisfied.

Subsequently, researchers in [13] conducted a study on time and cost optimization using the goal programming methods for port container handling to reduce the costs incurred and the length of time the queue optimization is performed on the scheduling at the port. The objective was to find the reason for which activity caused a delay in loading and unloading activities. From the result, the average was used as the target of the activity to be optimized and the maximum value as the input value to obtain the optimal value. The researcher determined that goal programming will be used to optimize time and costs for data with higher complexity and sample size in future work.

Next, [14] uses goal programming to optimize waste management. Some of the primary objectives (goals) are utilized interchangeably in this study to demonstrate the model's use as a tool in the planning process. The model is also a highly valuable tool for analyzing the cost-effectiveness of various priority orders, such as those led by environmental protection, cost minimization, energy recovery maximization, or resource recovery maximization. With the specified priority hierarchies, what is presented can be utilized to plan a long-term waste management system or to optimize waste management activities.

3. Methodology

3.1 Waste Sources Data

The study relies on data generously supplied by E – Idaman Sdn. Bhd, a private firm specializing in solid waste management, is responsible for overseeing the solid waste system in Northern Malaysia. The data, sourced specifically from their branch in Majlis Perbandaran (MP) Kangar, pertains to January 2023. There are three routes of data with two routes of collection and three types of vehicles used. The route obtained is the Beseri and Kangar to Jalan Santan route. These vehicles include the regular vehicle and the vehicle with an arm roll where the Kangar to Jalan Santan route performs collection using both normal and arm roll included vehicles. Arm roll is the machinery that can help lift bins and dumps without the need for manpower. Based on the data, five types of costs were obtained from each vehicle which are collection cost, labor cost, vehicle cost, consumable cost, and financial statement of operation. Table 1 shows the cost statement for E-Idaman Kangar in January 2023.

To break it down, the labor cost encompasses salaries for the management team, including the supervisor and manager, along with the wages for the driver and collection crew. Following that, the collection cost factors in the collection rate for each bin per route. Moving on, the vehicle cost encompasses maintenance, fuel consumption, road tax, and insurance. Lastly, consumable costs include expenses for tools such as scoops, rakes, and gloves used by the crew monthly.

Notably, the cost for the Kangar to Jalan Santan route with an arm roll is notably lower than the other routes, owing to reduced manpower requirements for collection due to the presence of an arm roll on the vehicle.

Table 1. Cost Statement in Ringgit Malaysia (RM) for E-Idaman Kangar, January 2023

Route	Labor Cost	Collection Cost	Vehicle Cost	Consumable Cost	Total
Beseri	16086.97	10114.97	6955.00	65.67	33222.61
Kangar – Jalan Santan	15558.17	12085.84	9685.00	65.67	37394.68
Kangar – Jalan Santan + arm roll	8937.36	7846.35	5057.50	52.67	21893.88
Total	40582.50	30047.16	21697.50	184.01	92511.17

3.2 Generate goal programming model.

In the following statement, decision variables, goals and constraints for the model are defined.

The decision variables are shown as follows:

x_j : the type of vehicles, j = 1,2,3 where;
x_1 = vehicle 1 (route Beseri)
x_2 = vehicle 2 (route Kangar – Jln Santan)
x_3 = vehicle 3 (route Kangar – Jln Santan + armroll)

The process of establishing and defining the objectives that a goal programming model seeks to accomplish is known as goal formulation. This study focuses on five goals which are outlined below:

Labor Cost Optimization

Minimize the salary expenses of labor.

$$\sum_{(i=1)}^{5} S_i x_j \leq T_S \tag{1}$$

$$\sum_{(i=1)}^{5} S_i x_j + d_1^- - d_1^+ = T_S \tag{2}$$

Where S_i is the salary paid for labor each month.
T_S is the target salary expenses of labor
d_1^- underachievement from the minimum labor cost per vehicle
d_1^+ overachievement from the minimum labor cost per vehicle

Collection Cost Optimization

Minimize the collection cost.

$$\sum_{(i=1)}^{5} C_i x_j \leq T_c \tag{3}$$

$$\sum_{(i=1)}^{5} C_i x_j + d_2^- - d_2^+ = T_C \tag{4}$$

Where C_i is the collection cost each month.

T_C is the target collection cost

d_2^- underachievement from the minimum collection cost per vehicle

d_2^+ overachievement from the minimum collection cost per vehicle

Vehicle Cost Optimization

Minimize the vehicle cost of the collection.

$$\sum_{(i=1)}^{5} V_i x_j \leq T_V \tag{5}$$

$$\sum_{(i=1)}^{5} V_i x_j + d_3^- - d_3^+ = T_V \tag{6}$$

Where V_i is the cost of vehicle collection each month.

T_V is the target vehicle cost for collection.

d_3^- underachievement from the minimum vehicle cost per vehicle

d_3^+ overachievement from the minimum vehicle cost per vehicle

Consumable Cost Optimization

Minimize the consumable cost.

$$\sum_{(i=1)}^{5} M_i x_j \leq T_M \tag{7}$$

$$\sum_{(i=1)}^{5} M_i x_{(j)} + d_4^- - d_4^+ = T_M \tag{8}$$

Where M_i is the consumable cost each month.

T_M is the target consumable cost.

d_4^- underachievement from the minimum consumable cost per vehicle

d_4^+ overachievement from the minimum consumable cost per vehicle

Financial Statement Optimization

Minimize the financial statement of operation.

$$\sum_{(i=1)}^{5} F_i x_j \leq T_F \tag{9}$$

$$\sum_{(i=1)}^{5} M_i x_j + d_5^- - d_5^+ = T_F \tag{10}$$

Where F_i is the financial statement of operation each month.

T_F is the target statement cost.

d_5^- underachievement from the minimum financial statement cost per vehicle

d_5^+ overachievement from the minimum financial statement per vehicle

Based on the goals derived from the equations (1) – (10) above, the corresponding achievement function is as below:

$$\text{Minimize} = P_1 d_1^+ + P_2 d_2^+ + P_3 d_3^+ + P_4 d_4^+ + P_5 d_5^+ \tag{11}$$

Where P_k : the priority coefficient for the kth priority level, k = 1,2,3,4,5

Lingo will employ the goal constraint to yield the desired outcomes. These achieved goal constraints are established based on the data set provided by E-Idaman Sdn. Bhd and the goal formulation derived from equations (2), (4), (6), (8), and (10).

4. Results and Discussion

4.1 Data Analysis
Upon executing the model in the Lingo software, the outcome of the model's performance is presented in the Lingo solver output as shown in Figure 1 and the comprehensive solution is detailed in the Lingo Optimization Report as shown in Figure 2. The Lingo solver furnishes the objective value achievable through the optimal solution. The Lingo solution report shows the detailed value of each deviation in achieving the goal and the value of the deviations that can be reduced.
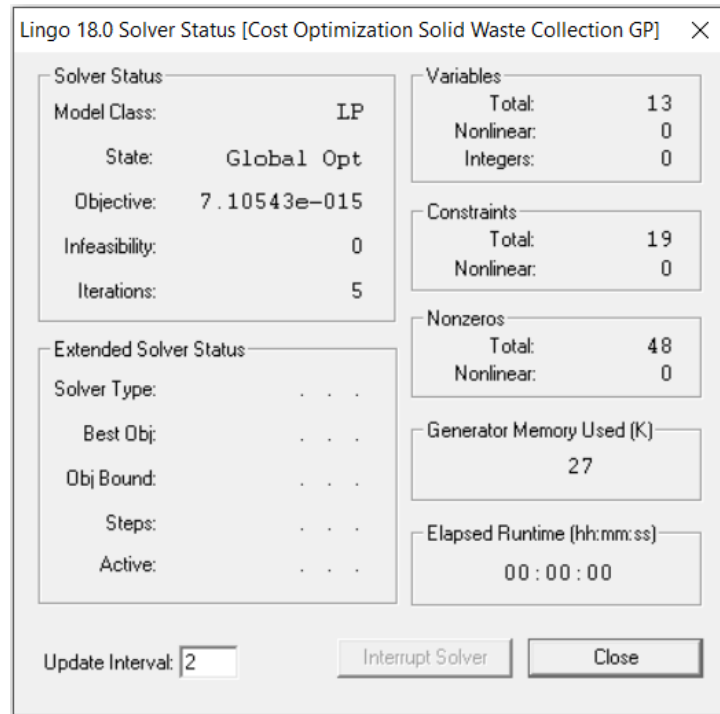
Figure 1: Lingo Solver Status
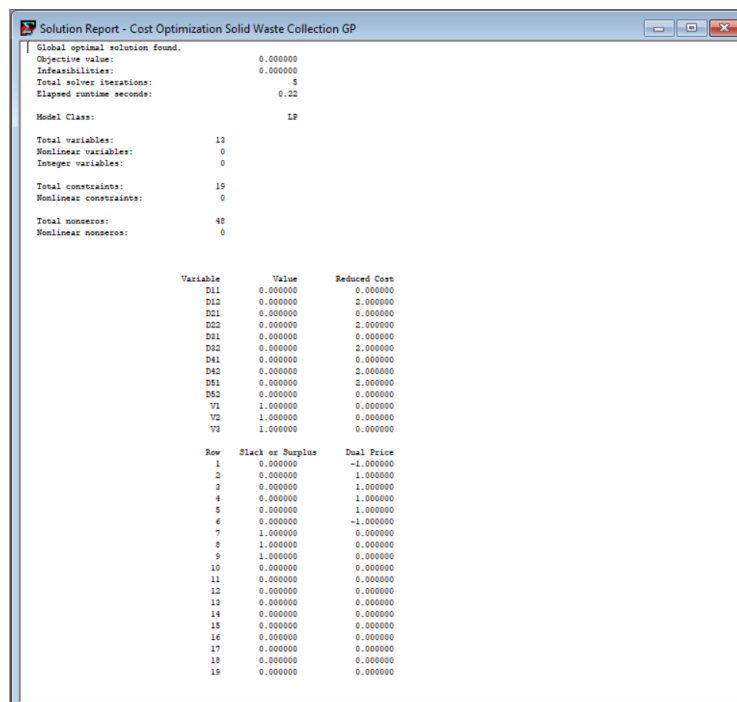
Figure 2. Lingo Optimization Report

From the Lingo Optimization Report in Figure 2, the result for goal achievement on the optimal solution was analyzed as presented in Table 2.

Table 2. Goals Achievement

Goals	Output value	Goals achievement
G1	$d_1^+=0$	Fully achieved
G2	$d_2^+=0$	Fully achieved
G3	$d_3^+=0$	Fully achieved
G4	$d_4^+=0$	Fully achieved
G5	$d_5^+=0$	Fully achieved

Based on findings illustrated in Figure 2, D12 is the deviation for labor cost to achieve the goal, where D22 is for collection cost, D32 is for vehicle cost, D42 is for consumable cost, and lastly, D51 is for financial statement operation. Based on the result shown in Table 2 for the five deviations, the total cost can be minimized when the cost of each vehicle is optimized.

Table 3. Results for deviations

Goals	d_i^+	d_i^-
G1	0	0
G2	0	0
G3	0	0
G4	0	0
G5	0	0

According to the data presented in Table 3, all goals (G1, G2, G3, G4, and G5) exhibit both positive and negative deviations, with $d_1^+=d_1^-=d_2^+=d_2^-=d_3^+=d_3^-=d_4^+=d_4^-=d_5^+=d_5^-=0$. This means that all the goals are fully achieved. Also, based on the negative deviations for all the goals, it shows that the cost is already at its optimal level. This indicates the full achievement of all goals. In other words, there is no further need to minimize the company's labor, collection, vehicle, consumable costs, and the financial statement of operation for solid waste collection, as they are already in their optimal form.

5. Conclusion

Malaysia is facing a growing problem with solid waste, largely due to the country's rapidly increasing population and rising living standards. Malaysia's current waste management system is insufficient to cope with the increasing amount of waste being produced. Out of the number of states in Malaysia, Perlis is also rapidly increasing waste collection, leading to an environmental issue. A private solid waste collection company called E-Idaman Sdn. Bhd. conveys the collection surrounding the whole Perlis state. This study focuses on optimizing the cost for cost reduction and optimizing the cost of solid waste collection by goal programming. Actual data was taken from the company, and three collection routes were analyzed with parameters of labor cost, collection cost, vehicle cost, and consumable cost. From this study, the hopes are to help with the future of financial and budget planning for the company to accomplish an optimal solid waste collection system. This could also create better decision-making and image for the company and the state.

The goal programming method can assist this company in achieving its optimal solution at a lower cost, according to the study's findings. The company has no requirement to reduce expenses since labor, collection, vehicle, and consumable costs are all currently optimized. All things considered, goal programming can assist in obtaining the best answers and cost optimization for the collection of solid waste. It is possible to conclude that the goals have been met from there. It is important to remember that goal programming is not a perfect solution, even with the results attained. It might not always be able to accomplish every objective at the same time.

References

[1]   Q. Zhuo and W. Yan, "Optimizing the number and location of household waste collection sites by multi-maximal covering location model: An empirical study in Minamata City, Kumamoto Prefecture, Japan," *J Clean Prod*, vol. 379, p. 134644, Dec. 2022, doi: 10.1016/j.jclepro.2022.134644.

[2]   Mohd Shahril, "Waste to Energy for A Sustainable Future - MIDA | Malaysian Investment Development Authority." Accessed: Feb. 27, 2024. [Online]. Available: https://www.mida.gov.my/waste-to-energy-for-a-sustainable-future/

[3]   M. S. Nikman Lee, S. Mohamed, M. A. Nasid Masrom, M. A. Abas, and S. T. Wee, "Risk in Green Retrofits Projects: A Preliminary Study on Energy Efficiency," *IOP Conf Ser Earth Environ Sci*, vol. 549, no. 1, p. 012084, Aug. 2020, doi: 10.1088/1755-1315/549/1/012084.

[4]   Z. Sakawi, "Municipal solid waste management in Malaysia: Solution for sustainable waste management," 2011. [Online]. Available: https://www.researchgate.net/publication/228406721

[5]   J. P. Malaysia, J. Perangkaan, M. Sedang Menjalankan, and S. Pendapatan, " Statistik Alam Sekitar Environment Statistics Perlis 2021 'Connecting the World with Data We Can Trust,'" May 2023. [Online]. Available: https://www.dosm.gov.my

[6]   C. Shiun Ming, C. Ngai Weng, G. Kai Chen, S. Ta Wee, K. Sie Long, and G. Hui Hwang, "A Review of Six Functional Elements Affect the Cost of Solid Waste Management in Malaysia," 2015.

[7]   A. Malakahmad, P. M. Bakri, M. R. M. Mokhtar, and N. Khalil, "Solid Waste Collection Routes Optimization via GIS Techniques in Ipoh City, Malaysia," *Procedia Eng*, vol. 77, pp. 20–27, 2014, doi: 10.1016/j.proeng.2014.07.023.

[8]   Y. Xu, S. Wu, H. Zang, and G. Hou, "An interval joint-probabilistic programming method for solid waste management: a case study for the city of Tianjin, China," *Front Environ Sci Eng*, vol. 8, no. 2, pp. 239–255, Apr. 2014, doi: 10.1007/s11783-013-0536-x.

[9]   Y. R. Fan, G. H. Huang, L. Jin, and M. Q. Suo, "Solid waste management under uncertainty: a generalized fuzzy linear programming approach," *Civil Engineering and Environmental Systems*, vol. 31, no. 4, pp. 331–346, Oct. 2014, doi: 10.1080/10286608.2014.913031.

[10]   H. Asefi, S. Shahparvari, and P. Chhetri, "Integrated Municipal Solid Waste Management under uncertainty: A tri-echelon city logistics and transportation context," *Sustain Cities Soc*, vol. 50, p. 101606, Oct. 2019, doi: 10.1016/j.scs.2019.101606.

[11]   A. Charnes and W. W. Cooper, "Goal programming and multiple objective optimizations," *Eur J Oper Res*, vol. 1, no. 1, pp. 39–54, Jan. 1977, doi: 10.1016/S0377-2217(77)81007-2.

[12]   H. A. Lyeme, A. Mushi, and Y. Nkansah-Gyekye, "Implementation of a goal programming model for solid waste management: a case study of Dar es Salaam – Tanzania," *International Journal for Simulation and Multidisciplinary Design Optimization*, vol. 8, p. A2, Jan. 2017, doi: 10.1051/smdo/2016018.

[13]   A. T. Rahman, R. Sarno, and Y. A. Effendi, "Goal programming to optimize time and cost for each activity in port container handling," in *2018 International Conference on Information and Communications Technology (ICOIACT)*, IEEE, Mar. 2018, pp. 866–871. doi: 10.1109/ICOIACT.2018.8350808.

[14]   A. V Shekdar and P. B. Mistry, "Evaluation of multifarious solid waste management systems - A goal programming approach," *Waste Management & Research: The Journal for a Sustainable Circular Economy*, vol. 19, no. 5, pp. 391–402, Oct. 2001, doi: 10.1177/0734242X0101900504.

.

Biography of all authors

	Diana Sirmayunie Mohd Nasir is a senior lecturer from the Universiti Teknologi MARA Perlis branch. She graduated with a master's degree in mathematics from Universiti Sains Malaysia. Her research interests include optimization and Mathematical Modelling.	Design and supervise research work, interpretation, and draft article.
	Salwani Asman with Bachelor of Science (Hons.) Management Mathematics from College of Computing, Informatics and Mathematics, Universiti Teknologi Mara Perlis Branch.	Data analysis, interpretation and data collection of waste collection at E – Idaman Sdn. Bhd,Perlis.
	Suzanawati Abu Hasan is a senior lecturer at the Universiti Teknologi MARA Perlis campus. She holds a Master of Mathematics degree from Universiti Sains Malaysia. Her research interests include Applied Mathematics, Fuzzy Logic, and Optimization.	Literature review and camera ready.

Total Variation Selective Segmentation-based Active Contour Model with Distance Function and Local Image Fitting Energy for Medical Images

Nadiah Mohamed
School of Mathematical Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Negeri Sembilan Campus, Malaysia
nadiah@uitm.edu.my

Abdul Kadir Jumaat
School of Mathematical Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Malaysia
Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia
abdulkadir@tmsk.uitm.edu.my

Rozi Mahmud
Radiology Department, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang, Selangor, Malaysia
rozi@upm.edu.my

Article Info

ABSTRACT

The Active Contour Model (ACM) is a mathematical model in image processing that is commonly utilized to partition or segment an image into specific objects. The segmentation method in region-based ACM can be categorized into two classes: global ACM and selective ACM Selective ACM isolates a specific target item from an input image, which is more advantageous than the global ACM due to its proven use, particularly in medical image analysis. However, the selective ACM appears to produce poor outcomes when segmenting an image with uneven (inhomogeneous) intensity. Additionally, the current selective ACM that uses the Gaussian function as a regularizer generates a non-smooth segmentation curve, especially for images containing noise. This study introduces a new selective ACM that is designed to segment medical images with inhomogeneous intensity levels. The model incorporates a Total Variation term as a regularizer, distance function, and local image fitting concepts. The Euler-Lagrange (EL) equation was given to solve the suggested model, which is approximately 5% more accurate with a processing time that is around three times faster than the existing model, as shown by numerical testing. The suggested mathematical model can be advantageous for the image analysis community, particularly in the medical industry, to automatically segment a specific object in a medical image.

Corresponding Author:
Abdul Kadir Jumaat
School of Mathematical Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Malaysia
Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Universiti Teknologi MARA, 40450, Shah Alam, Selangor, Malaysia
Email: abdulkadir@tmsk.uitm.edu.my

1. Introduction

In image processing field, the image segmentation approach either variational or non-variational approaches is an essential step involves in separating a digital picture into different portions for further investigations in many application such as recognition of object, medical image analysis and computer vision [1-5]. Non-variational approaches such as thresholding and region

growth are good for basic images but struggle with topological changes. Another well-known non-variational method namely the learning methods (machine or deep learning) demand a significant quantity of labelled data, which is often not accessible. In contrast, variational methods are less reliant on the quantity of data and additionally they are demonstrated efficacy in picture segmentation.

In variational approaches, the Active contour model (ACM) is a popular method that utilize the calculus of variations to apply optimization strategies to minimize the energy cost function. There are two forms of variational ACM: edge-based and region-based methods. A prominent edge-based variational ACM namely Snake model [6] was formulated in explicit parametric contour. The drawback of this technique is susceptible to picture noise. In contrast, the region-based variational ACM is less affected by image noise and capable to manage topological changing as the method considers statistical features of an input images in the formulation [7-9]. This region-based variational ACM may be categorized into two classes: global ACM and selective ACM.

The most recognized variational global active contour model was the Mumford-Shah model developed in [10]. Being a global type of model, its objective is to segment all items or objects included in the input picture. A simplified version [11] of the Mumford-Shah model into a precise numerical representation known as the Chan-Vese (CV) model. However, the CV model may yield unsatisfactory outcomes as this model's reliance on global image intensity information in its mathematical formulation hinders its ability to effectively segment images that have intensity inhomogeneity.

Intensity inhomogeneity is a significant limitation in image segmentation, especially in medical imaging due to presence of noise or problem with the imaging modalities. This pertains to fluctuations in picture intensity levels within a single image, causing challenges, for instance, in distinguishing between healthy and abnormal tissue that results in erroneous judgements among medical professionals. This may be detrimental, as for disorders like cancer, precise diagnosis is crucial for planning efficient therapy.

2. Literature Review

The Local Binary Fitting (LBF) model [12] was presented to address the issue of segmenting an image with intensity inhomogeneity by including local image intensity. However, the accurate segmentation result achieved by using the LBF model led to longer processing time due to its high computational complexity. Therefore, another ACM model namely Local Image Fitting (LIF) model was proposed [13] to address the shortcomings of the LBF approach. Subsequently, several studies have applied similar concept of LIF model as demonstrated by [14-16]. However, the LIF model cannot be used to segment a particular item in an input picture.

To address the issue of segmenting a particular item in an input picture, the selective segmentation model (selective ACM), which is the second type of variational region-based ACM, is recommended. Selective segmentation is the process of partitioning a particular item in an image into multiple segments based on markers specified by the user. This technique shows promise for integration with medical imaging [17-[8], biometric identification [19], and text analysis [20]. A selective ACM algorithm called Primal Dual Selective Segmentation (PDSS) was introduced in [21]. This approach has been demonstrated to be effective at isolating a particular item within an input image. However, the PDSS model encounters a significant challenge in segmenting pictures with low contrast. Hence, a modified PDSS model called PDSS2 [22], which replaces the fitting term with information obtained from the image enhancement approach to address the issue. Unfortunately, the precise information obtained by the PDSS2 appears inadequate for identifying the near-optimal segmented border of an object with intensity inhomogeneity issues.

Recently, a novel variational selective region-based ACM was introduced in 2023 called the Gaussian regularization selective segmentation (GRSS) model [23] for pictures with intensity inhomogeneity. However, the curve created using a Gaussian kernel regularizer in the formulation was not smooth because the regularizer is sensitive to image noise. Additionally, it is difficult to tune the parameter that control the regularizer. Besides Gaussian kernal function, another more powerful regularizer is to utilize the Total variation (TV) functional. The CV model [11] was an early mathematical model that incorporated TV in its formulation, leading to a smooth segmentation curve. However, as the model was formed inside a variational global region-based segmentation framework, it was not able to effectively segment a specific item in an image.

This study introduces a novel variational selective region-based ACM called the Total Variation based Selective Segmentation (TVSS) model for pictures with intensity inhomogeneity that

incorporates the concept of employing TV as a regularizer. In addition, the local attribute of an input picture was included using the concept from the LIF model [13]. A distance function from the GRSS model [23] was used to capture a specific target.

This article is divided into three portions. Section 3 outlines the methodology of the study and introduces the suggested energy function of the proposed model. Section 4 includes a comparison and explanation of the experimental results. Section 5 discusses the findings and recommendations for further research.

3. Methodology

This section outlines the methodology of the study, as depicted in the flow chart shown in Figure 1.
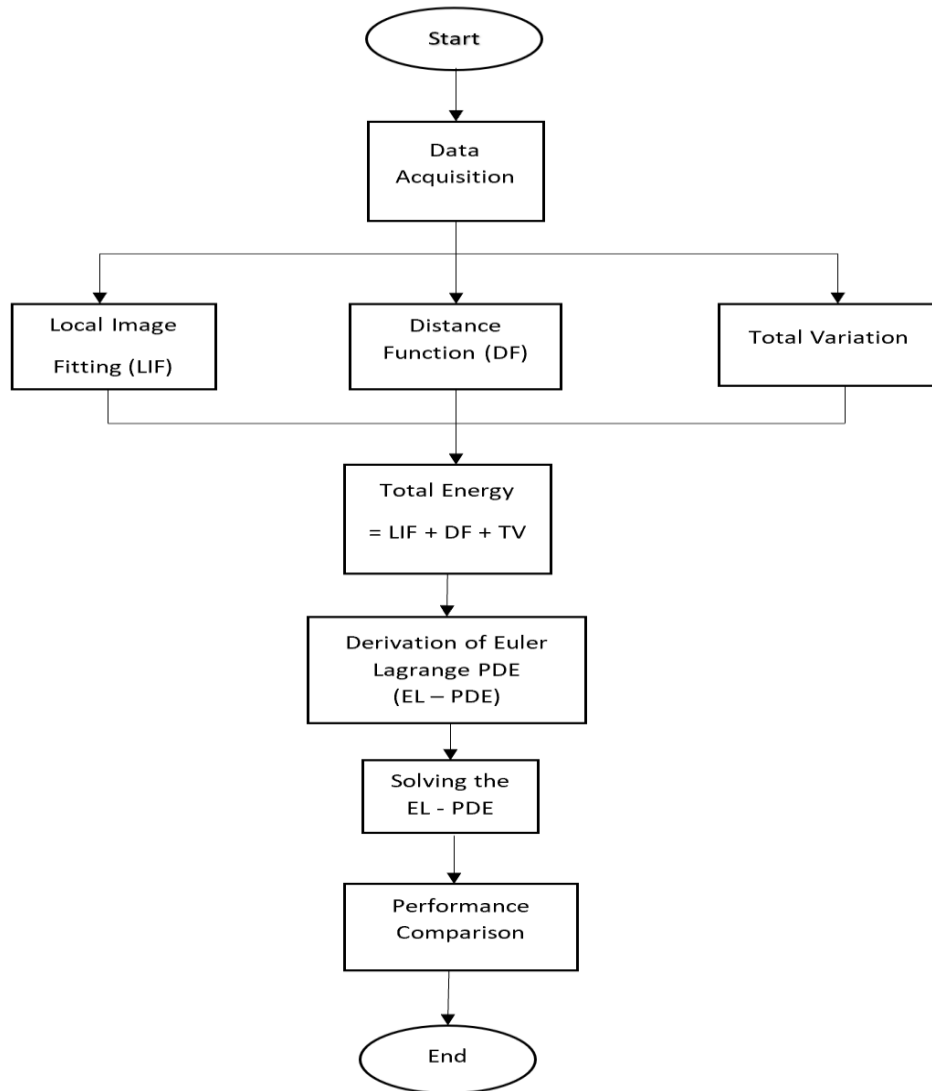
Figure 1. Flowchart of the research methodology

Figure 1 depicts the flowchart outlining the research methodology used in this study. The Local Image Fitting (LIF) energy, Distance Function (DF), and TV of an input picture were calculated at the beginning. A total energy minimization function was created to reflect the suggested TVSS model to segment the input picture. This function integrates the LIF energies, DF, and TV term. The Euler-Lagrange Partial Differential Equation (EL-PDE) of the TVSS model was generated and solved. The performance of the proposed TVSS model was then compared with the GRSS model. In the following section, each stage is described in detail.

3.1 Data Acquisition

In this study, six (6) sets of mammography test images and six (6) sets of chest X-ray test images with their ground truth segmentation solutions were collected from [24] and [25] respectively. The image dimension of 256 by 256 pixels was utilized for all test images. In addition, the test images had varying intensities to ensure that the proposed model met its objective in segmenting images with intensity inhomogeneity.

3.2 Local Image Fitting (LIF)

The LIF, which is important to handle images with intensity inhomogeneity, was computed using the LIF model [13]. Assuming that for an image $I = I(x, y)$ in a domain Ω, the LIF energy function in the level set formulation, $E_{\varepsilon}^{LIF}(\phi)$ is defined as follows:

$$E_{\varepsilon}^{LIF}(\phi) = \frac{1}{2}\int_{\Omega}\left|I - I^{LIF}\right|^2 dxdy \tag{1}$$

Here, the function $\phi(x, y)$ is a level set function that represents the segmentation contour. The fitted image $I^{LIF} = I^{LIF}(x, y)$ in equation (1) can be written as $I^{LIF} = n_1 H(\phi) + n_2(1 - H(\phi))$ where $H(\phi)$ is a heaviside function or also known as the characteristics function. The value $H(\phi)$ is 1 inside the contour and 0 outside the contour. Next, $n_1 = \text{mean}\left(I \in \left(\{(x, y) \in \Omega | \phi < 0\} \cap W_k\right)\right)$ and $n_2 = \text{mean}\left(I \in \left(\{(x, y) \in \Omega | \phi < 0\} \cap W_k\right)\right)$ are the intensity averages of interior and exterior in a local region respectively.

3.3 Distance Function (DF)

In the GRSS model [23], the marker set was introduced, and it was defined as $A = \{w_i = (x_i^*, y_i^*) \in \Omega, 1 \le i \le n_1\}$ with $n_1(\ge 3)$ marker points that were placed near the targeted object. Then, the DF which is the Euclidean distance of each point $(x, y) \in \Omega$ from its nearest point in the polygon, P made up of $(x_p, y_p) \in P$, constructed from the user input set, A was defined as follows:

$$P_d(x, y) = \sqrt{\left(x - x_p\right)^2 + \left(y - y_p\right)^2} \tag{2}$$

Here, the DF was introduced as a penalty term so that the evolving segmentation contour remains close to the targeted object indicated by the marker set.

3.4 Total Variation (TV) Term

In modeling an ACM, the TV term is used to regularize the generated segmentation curve. One of the earliest ACM that utilized the TV term is the CV model [11]. The TV term was defined as follows:

$$TV = \int_{\Omega}\delta(\phi)\left|\nabla\phi\right| dxdy \tag{3}$$

Here, the function $\delta(\phi)$ is the Dirac delta functional which is equivalent to $H'(\phi)$ and the function $\left|\nabla\phi\right|$ is the magnitude of gradient ϕ. By minimizing the TV term, the length of the generated curve was optimized to ensure it is short and smooth. Although the CV model applies to the TV term, the generated segmentation curve by the model was over-segmented especially in segmenting one object among other objects. This was mainly because the model is a global type of segmentation model.

3.5 Total Energy Function

The proposed TVSS model was formulated by modeling a total energy minimization function that utilized the LIF energy, DF, and TV term. As a result, the proposed TVSS model was defined as follows:

$$\min_{\phi} \{ \ TVSS(\phi) = v\int_{\Omega} \delta(\phi)|\nabla\phi| \ dxdy + \frac{1}{2}\int_{\Omega} \left(I - \left(n_1 H(\phi) + n_2 \left(1 - H(\phi)\right)\right)\right)^2 dxdy$$
$$+ \theta\int_{\Omega} P_d H(\phi) \ dxdy \ \} \tag{4}$$

with $n_1(x,y) = k_{\sigma}*\left[H(\phi)I\right]/k_{\sigma}*H(\phi)$ and $n_2(x,y) = k_{\sigma}*[1-H(\phi)]I/k_{\sigma}*[1-H(\phi)]$. The function k_{σ} is a Gaussian kernel with standard deviation σ such that $k_{\sigma} = e^{-(x^2+y^2)/2\sigma^2}$. The parameter θ restricts the contour from evolving too far from the targeted object. Normally, smaller θ is suitable if the targeted object can be clearly distinguished form the background and vice versa [26-27]. To obtain a smooth contour, the TV term which is the first integrand was used to regularize the generated segmentation curve. The term was weight by the parameter v. For noisy images, a large v can be imposed.

3.6 Derivation of Euler Lagrange Partial Differential Equation (EL-PDE)

The proposed TVSS model in equation (4) was minimized using calculus of variation by deriving its EL-PDE. Firstly, n_1 and n_2 was kept fixed and minimized the equation (4) with respect to ϕ using the Gateaux derivative. In calculus of variation, the Gateaux derivative is used to find the first variation of the model formulation function with respect to ϕ. The Gateaux derivative of $TVSS$ at a point ϕ and a test function ψ, denoted by $\frac{\partial}{\partial\phi}TVSS(\phi,\psi)$, is defined as the limit which is illustrated in the following equation (5).

$$\frac{\partial}{\partial\phi}TVSS(\phi,\psi) = \lim_{h\to 0}\frac{TVSS(\phi+h\psi)-TVSS(\phi)}{h} = \frac{d}{dh}TVSS(\phi+h\psi)\Big|_{h=0} \tag{5}$$

Thus, by applying the equation (5), we arrived at the following equation (6):

$$\frac{d}{dh}\int_{\Omega}\frac{1}{2}\left[I-\left(n_1 H(\phi+h\psi)+n_2\left(1-H(\phi+h\psi)\right)\right)\right]^2 dxdy\Big|_{h=0}$$
$$+\frac{d}{dh}\int_{\Omega}\theta P_d H(\phi+h\psi) \ dxdy\Big|_{h=0} + \frac{d}{dh}v\int_{\Omega}\delta(\phi+h\psi)|\nabla(\phi+h\psi)| \ dxdy\Big|_{h=0} \tag{6}$$

Differentiating all the integrands involved in equation (6) yielded:

$$\int_{\Omega}\left[I-\left(n_1 H(\phi+h\psi)\right)+n_2\left(1-H(\phi+h\psi)\right)\right]\left[-n_1 H'(\phi+h\psi)\psi+n_2 H'(\phi+h\psi)\psi\right]d\Omega\Big|_{h=0}$$
$$+\int_{\Omega}v\delta(\phi+h\psi)\frac{d}{dh}\left(|\nabla(\phi+h\psi)|\right)d\Omega\Big|_{h=0} + \int_{\Omega}|\nabla(\phi+h\psi)|v\frac{d}{dh}\delta(\phi+h\psi)d\Omega\Big|_{h=0}$$
$$+\int_{\Omega}\theta P_d H'(\phi+h\psi)\psi d\Omega\Big|_{h=0}$$

After simplification, we obtained the following equation (7) as follows:

$$\int_{\Omega}\delta(\phi)\psi\left[I-\left(n_1 H(\phi)+n_2\left(1-H(\phi)\right)\right)\right]\left[-n_1+n_2\right]dxdy+\int_{\Omega}v\delta(\phi)\frac{\nabla\phi\cdot\nabla\psi}{|\nabla\phi|}dxdy$$
$$+\int_{\Omega}v|\nabla\phi|\delta'(\phi)\psi dxdy+\int_{\Omega}\theta P_d\delta(\phi)\psi dxdy = 0 \tag{7}$$

Green's theorem was applied to further simplify the equation (7) that yielded the following equation (8):

$$\int_{\Omega}\delta(\phi)\psi\left[I-\left(n_1 H(\phi)+n_2\left(1-H(\phi)\right)\right)\right]\left[-n_1+n_2\right]dxdy -$$
$$\int_{\Omega}v\psi\delta(\phi)\nabla\cdot\frac{\nabla\phi}{|\nabla\phi|}dxdy + \int_{\partial\Omega}v\psi\frac{\delta(\phi)}{|\nabla\phi|}\cdot\frac{\partial\phi}{\partial n}dS + \int_{\Omega}\theta P_d\delta(\phi)\psi dxdy \tag{8}$$

Here, n is the unit normal vector and S is the arc length. Note that the EL-PDE is defined when $\frac{\partial}{\partial\phi}TVSS(\phi,\psi)=0$. Thus, for all test function ψ, the integrands in equation (8) will be zero if

$$\delta\left(\phi\right)\Big[I-\left(n_1 H\left(\phi\right)+n_2\left(1-H\left(\phi\right)\right)\right)\Big]\left[-n_1+n_2\right]-\nu\delta\left(\phi\right)\nabla\cdot\frac{\nabla\phi}{\left|\nabla\phi\right|}+\theta P_d\delta\left(\phi\right) \tag{9}$$

with Neumann Boundary condition such that $\dfrac{\partial\phi}{\partial n}\left[\nu\,\dfrac{\delta\left(\phi\right)}{\left|\nabla\phi\right|}\right]=0\Rightarrow\dfrac{\partial\phi}{\partial n}=0$. Here, equation (9) is called

the EL-PDE of the proposed TVSS model.

3.7 Solving the EL-PDE of the Proposed TVSS Model

To solve the EL-PDE in equation (9) iteratively, the gradient descent method is applied where an artificial time step t is introduced such that

$$\begin{aligned}\frac{\partial\phi}{\partial t}&=-\frac{\partial}{\partial\varphi}TVSS\\&=\delta\left(\phi\right)\Big[I-\left(n_1 H\left(\phi\right)+n_2\left(1-H\left(\phi\right)\right)\right)\left(n_1-n_2\right)\Big]+\nu\delta\left(\phi\right)\nabla\cdot\frac{\nabla\phi}{\left|\nabla\phi\right|}-\theta\delta\left(\phi\right)P_d\end{aligned} \tag{10}$$

Equation (10) was discretized and solved using the explicit finite difference scheme using the following equation (11).

$$\begin{aligned}\phi_{i,j}^{n+1}&=\phi_{i,j}^{n}+\Delta t\,\Big\{\ \delta\left(\phi_{i,j}^{n}\right)\Big[I-\left(n_1 H\left(\phi_{i,j}^{n}\right)+n_2\left(1-H\left(\phi_{i,j}^{n}\right)\right)\right)\left(n_1-n_2\right)\Big]\\&\quad+\nu\delta\left(\phi_{i,j}^{n}\right)\nabla\cdot\frac{\nabla\phi_{i,j}^{n}}{\left|\nabla\phi_{i,j}^{n}\right|}-\delta\left(\phi_{i,j}^{n}\right)\left[\theta P_d\right]\ \Big\}\end{aligned} \tag{11}$$

3.8 Performance Comparison

The segmentation performance was judged based on the quantitative accuracy measures using the Dice Similarity Coefficient (DSC) and Jaccard Similarity Coefficient (JSC) metrics. The metrics were computed based on the following formulae:

$$\mathrm{JSC}=\frac{|S_n\cap S_*|}{|S_n\cup S_*|},\ \ \mathrm{DSC}=\frac{|S_n\cap S_*|}{|S_n|+|S_*|} \tag{12}$$

where S_n is the set of segmentation domain generated by GRSS or TVSS model and S_* is the ground truth solution domain. The return value of DSC and JSC that approaches 1 indicates good quality segmentation. The value that approaches 0 indicates poor segmentation quality. The processing time was also recorded to determine the efficiency of the GRSS model and the TVSS model.

3.9 TVSS Model's Algorithm

To compute the solution of equation (11), the MATLAB R2017b software was used in a laptop with the specification of Processor: Intel(R) Core(TM) i7- 6500 CPU @ 2.50GHz 2.60 GHz installed memory (RAM): 8 GB, System type: 64 - boy operating system, 64-based processor. The following Figure present the algorithm to implement the TVSS model.

```
1.  Input the test image.
        >> Img = imread('image.png');
2.  Set values of tol, maxit, σ and θ.
        >> maxit = 100; tol=1e-6; theta=20; sigma =90;
3.  Marker set is defined around the targeted object.
        >> mx = [125;62;97;158;214;174]; my = [56;120;147;161;150;62];
4.  For iteration=1 to maxit or ‖φⁿ⁺¹ − φⁿ‖/‖φⁿ‖ ≤ tol do
    Generate the function φ which is the zero-level set function based on Equation (11).
        >> for iteration = 1: maxit
        >> [phi]= TVSS(Img,mx,my,theta,sigma,maxit,tol);
        >> R=Residual(phi,oldphi)/norm(oldphi);
        >> if R<tol, break, end
        >> end
5.  The final segmentation curve will be the output φ.
        >> figure, imagesc(phi);colormap gray;
```

Figure 1. Steps to implement the TVSS model

4. Results and Discussion

Figure 2 shows all 12 sets of test images with the initial contour in yellow and marker set in green used in this experiment.

(a) Test 1	(b) Test 2	(c) Test 3	(d) Test 4
(e) Test 5	(f) Test 6	(g) Test 7	(h) Test 8
(i) Test 9	(j) Test 10	(k) Test 11	(l) Test 12
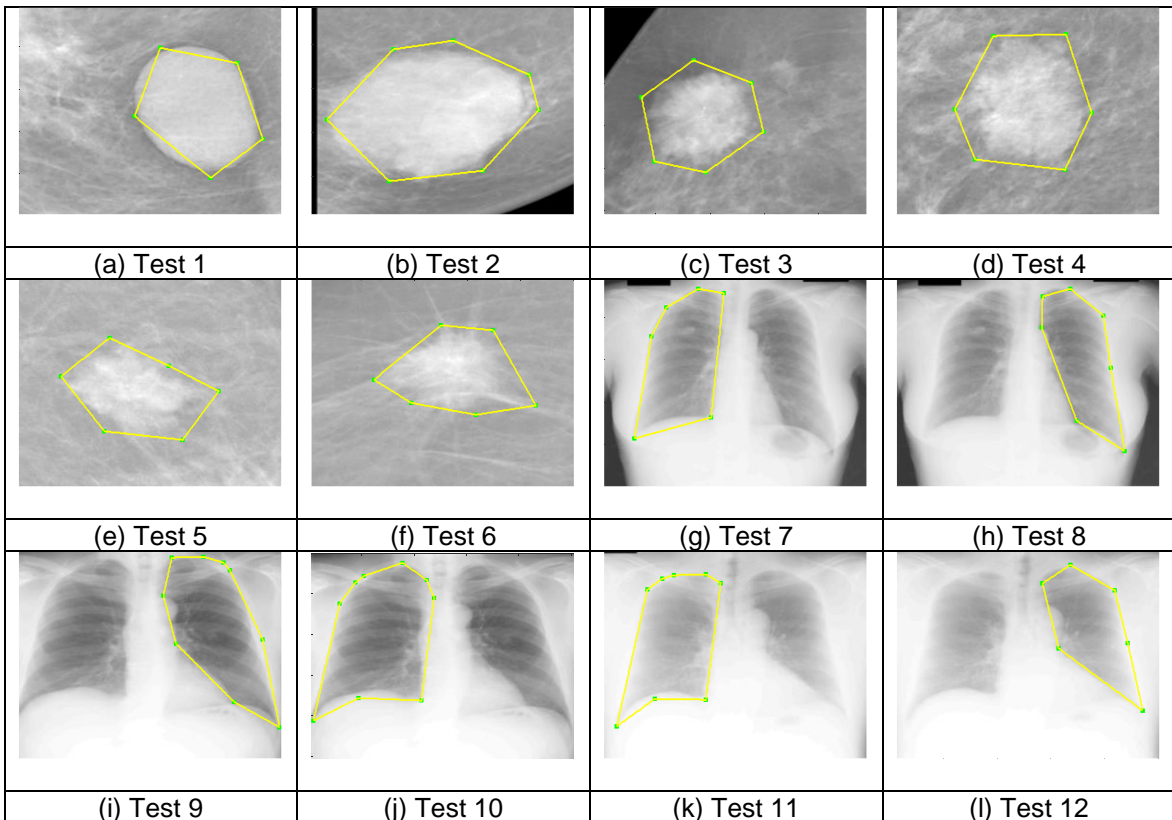
Figure 2. Test images with initial contours and marker set

From Figure 2, Test 1 until Test 6 were mammography images, while Test 7 until Test 12 were chest x-ray images. These types of images were chosen because all the test images were challenging to be segmented due to low contrast appearance. Additionally, the images had inhomogeneous intensity. Hence, all the images chosen were significant in testing the performance of the proposed model. The segmentation process was run using the MATLAB software, and the

segmentation results of all test images were compared between the proposed TVSS model and the GRSS model [24].

4.1 Segmentation of Mammography Images

The visual illustration of the segmentation result for the mammography images are demonstrated in the following Figure 3.
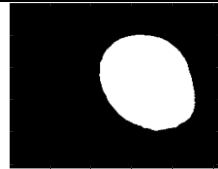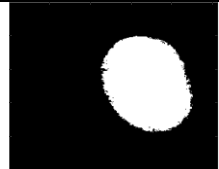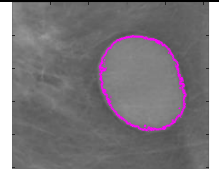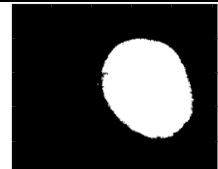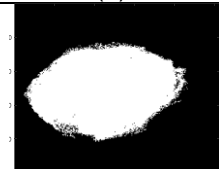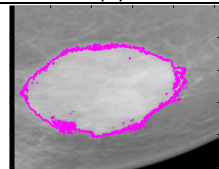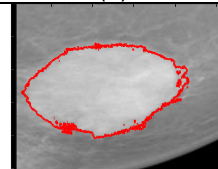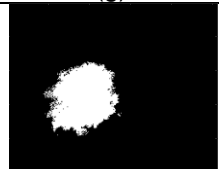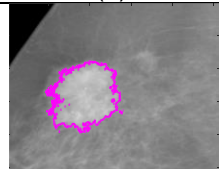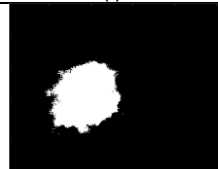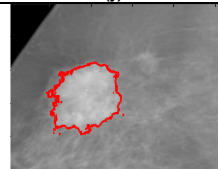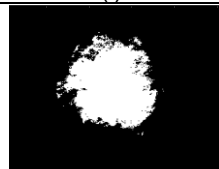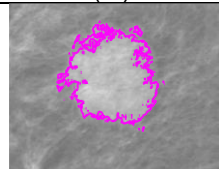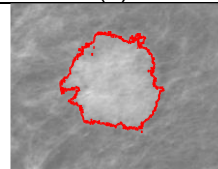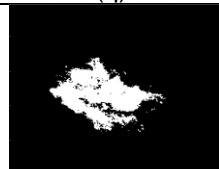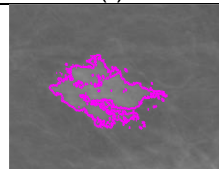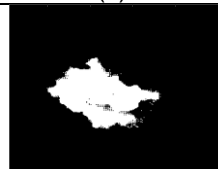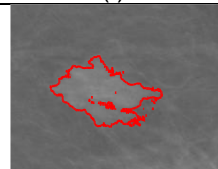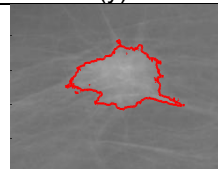
Benchmark	GRSS		TVSS	
(a) Test 1	(b)	(c)	(d)	(e)
(f) Test 2	(g)	(h)	(i)	(j)
(k) Test 3	(l)	(m)	(n)	(o)
(p) Test 4	(q)	(r)	(s)	(t)
(u) Test 5	(v)	(w)	(x)	(y)
(z) Test 6	(aa)	(ab)	(ac)	(ad)

Figure 3. Segmentation results from GRSS model and TVSS model in segmenting mammography images

As shown in Figure 3, the benchmarks (ground truth solutions) of the test mammography images are shown in the first column. There are two types of results generated from each model. The first type is the binary form as shown in the second and fourth columns which represent the results by the GRSS model and TVSS model respectively. The second type is the contour or curve representation as demonstrated in the third and last columns which demonstrate the results generated by GRSS model and TVSS model respectively.

By visual observation, both models were able to segment the targeted region (breast abnormalities) of all the test images. However, both produced different segmentation results. For instance, it is clear that the segmentation results for the GRSS model in Figure 3 (m, r and w) produced large amounts of unnecessary artifacts, which made the final contour more scattered. On the other hand, the result produced by the proposed TVSS was smoother as indicated in Figure 3 (o, t and y). These results demonstrate the advantage of using the TV term in the proposed TVSS model compared to the Gaussian regularization term utilized in the GRSS model. By incorporating the TV term in the TVSS model, the length of the segmentation curve was minimized that gave a smoothing effect to the final segmentation curve.

Besides the qualitative visual observation, this experiment was also judged based on the quantitative accuracy measures using the Dice Similarity Coefficient (DSC) and Jaccard Similarity Coefficient (JSC) metrics. Additionally, the processing time was also recorded to determine the efficiency of the GRSS model and the TVSS model. Table 1 demonstrates the JSC, DSC and processing time of both models in segmenting all mammography test images.

Table 1. JSC, DSC and Processing Time for Segmenting Mammography Images

Image	JSC		DSC		Processing Time	
	GRSS	TVSS	GRSS	TVSS	GRSS	TVSS
Test 1	0.8944	0.9335	0.9443	0.9656	38.0244	9.4841
Test 2	0.8976	0.9032	0.9460	0.9491	71.8910	83.9554
Test 3	0.7779	0.7946	0.8751	0.8856	21.9318	22.3152
Test 4	0.7383	0.8144	0.8495	0.8977	137.3918	136.3861
Test 5	0.7811	0.8845	0.8771	0.9387	11.3808	11.6315
Test 6	0.9132	0.9242	0.9547	0.9606	126.2228	127.4576
Average	0.8338	**0.8757**	0.9078	**0.9329**	67.8071	**65.2050**

From Table 1, by taking the average of the JSC and DSC values for both models in segmenting the object's boundaries in all mammography test images, it was observed that the average JSC and DSC values for the proposed model were 0.8757 and 0.9329, respectively, which were 4.78% and 2.69% higher than the average JSC and DSC values recorded in the GRSS model. It was also observed that the processing time for both models to obtain the final segmentation result was comparable.

4.2 Segmentation of Chest X-ray Images
Next, the performance of GRSS and the proposed TVSS models in segmenting the chest x-ray images were compared. The results are illustrated in Figure 4. A similar observation as in the previous experiment in segmenting mammography images was observed where the segmentation curves generated by the proposed TVSS model were smoother compared to the GRSS model. This was mainly because the proposed model utilized the TV term which was capable of regularizing contours effectively compared to the Gaussian term applied in the GRSS model. The results were also compared quantitively by measuring their JSC, DSC and processing time as tabulated in Table 2.
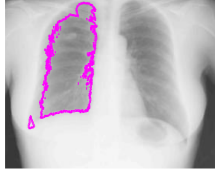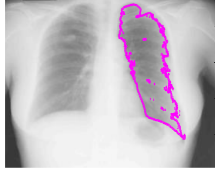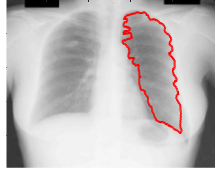
Benchmark	GRSS		TVSS	
(a) Test 7	(b)	(c)	(d)	(e)
(f) Test 8	(g)	(h)	(i)	(j)
(k) Test 9	(l)	(m)	(n)	(o)
(p) Test 10	(q)	(r)	(s)	(t)
(u) Test 11	(v)	(w)	(x)	(y)
(z) Test 12	(aa)	(ab)	(ac)	(ad)

Figure 4. Segmentation results from GRSS model and TVSS model in segmenting chest X-ray images

Table 2. JSC, DSC and Processing Time for Segmenting Chest X-ray Images

Image	JSC		DSC		Processing Time	
	GRSS	TVSS	GRSS	TVSS	GRSS	TVSS
Test 7	0.7108	0.8355	0.8310	0.9104	16.5518	5.9613
Test 8	0.7507	0.7766	0.8576	0.8743	8.7095	6.0210
Test 9	0.7455	0.8650	0.8542	0.9276	16.6471	5.9322
Test 10	0.7958	0.9250	0.8863	0.9610	16.7538	5.8816
Test 11	0.7247	0.8751	0.8404	0.9334	16.7853	5.8733
Test 12	0.7013	0.8755	0.8244	0.9336	17.3944	5.8171
Average	0.7381	**0.8588**	0.8490	**0.9234**	15.4737	**5.9144**

As depicted in Table 2, By taking the average of the JSC and DSC values for both models in segmenting the object's boundaries in all chest X-ray test images, certain patterns were observed. The average JSC and DSC values for the proposed model were 0.8588 and 0.9234, respectively. These values were 14.05% and 8.06% higher than the average JSC and DSC values recorded in the GRSS model. It was also observed that the proposed model was more efficient than the GRSS model, as it took less time to obtain the final segmentation result as compared to the GRSS model.

5. Conclusion

A new variational selective region-based ACM called the TVSS model was developed in this study. It combined the TV term with local image intensities. The EL equation was derived to ensure optimality and was subsequently solved using MATLAB. The TVSS model was compared to the GRSS model in terms of segmentation accuracy and efficiency. Segmentation accuracy was assessed using the average JSC and DSC values, while processing time was recorded to evaluate segmentation efficiency. All models successfully segmented the specified items visually, as indicated by the findings. However, the TVSS model was able to generate a smoother segmentation curve compared to the GRSS model. Additionally, the proposed TVSS model had the highest accuracy with a comparable processing time. The proposed TVSS model can be potentially commercialized by developing a software for image analysis that serves as a second eye to physicians or radiologists in interpreting a medical image for further decision. To enhance accuracy, the concept of saliency from references [28-29] can be included in future work.

Conflict of Interest
The authors declare no conflict of interest in the subject matter or materials discussed in this manuscript.

References

[1] S. N. Kumar, A. Lenin Fred, H. Ajay Kumar, and P. Sebastin Varghese, "Performance metric evaluation of segmentation algorithms for gold standard medical images," in *Adv. Intell. Syst. and Comput.*, vol. 709, 2018, pp. 457–469, doi: 10.1007/978-981-10-8633-5_45.

[2] N. Ismail, A. K. Jumaat, and N. F. A. Zulkarnain, "An Improved Variational-Based Model for Denoising and Segmentation of Vector-Valued Images," *J. Adv. Res. Appl. Sci. Eng. Tech.*, vol. 40, no. 1, pp. 189–203, 2024.

[3] M. S. Mazlin, A. K. Jumaat, and R. Embong, "Partitioning intensity inhomogeneity colour images via Saliency-based active contour," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 1, pp.337–346, 2024.

[4] N. A. Kon, A. K. Jumaat, and M. D. A. Suhaizi, "Active Contour Models for Boundary Extraction with Application to Medical Images with Noise," *J. Adv. Res. Appl. Sci. Eng. Tech.*, vol. 33, no. 2, pp. 300–312, 2024.

[5] M. S. Mazlin, A. K. Jumaat, and R. Embong, "Saliency-based variational active contour model

for image with intensity inhomogeneity," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 32, no. 1, pp.206–215, 2023.

[6] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. comput. vision*, vol. 1, pp. 321-331, 1988, doi:10.1007/BF00133570.

[7] S. Osher and J. A. Sethian, "Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations," *J. comput. phys.*, vol. 79, pp. 12-49, 1988, doi:10.1016/0021-9991(88)90002-2.

[8] H. Ali, L. Rada and N. Badshah, "Image Segmentation for Intensity Inhomogeneity in Presence of High Noise," *IEEE Transactions on Image Processing*, vol. 27, pp. 3729-3738, 2018, doi:10.1109/TIP.2018.2825101.

[9] S. Soomro, A. Munir and K. N. Choi, "Fuzzy c-means clustering based active contour model driven by edge scaled region information," *Expert Syst. with Appl.*, vol. 120, pp. 387-396, 2019, doi:10.1016/j.eswa.2018.10.052.

[10] D. B. Mumford and J. Shah, "Optimal approximations by piecewise smooth functions and associated variational problems," *Commun. Pure and Appl. Math.*, vol. 42, pp. 577-685, 1989, doi:10.1002/cpa.3160420503.

[11] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, pp. 266-277, 2001, doi:10.1109/83.902291.

[12] C. Li, C. Y. Kao, J. C. Gore and Z. Ding, "Implicit Active Contours Driven by Local Binary Fitting Energy," In *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, 2007, pp. 1-7, doi: 10.1109/CVPR.2007.383014.

[13] K. Zhang, H. Song and L. Zhang, "Active contours driven by local image fitting energy," *Pattern Recognit.*, vol. 43, pp. 1199-1206, 2010, doi:10.1016/j.patcog.2009.10.010.

[14] H. Wang, and T. Z. Huang, "An adaptive weighting parameter estimation between local and global intensity fitting energy for image segmentation," *Commun in Nonlinear Sci. and Numer. Simul.*, vol. 19, pp. 3098-3105, 2014, doi:10.1016/j.cnsns.2014.02.015.

[15] Y. Yang, W. Jia and B. Wu, "Simultaneous segmentation and correction model for color medical and natural images with intensity inhomogeneity," *The Visual Comput.*, vol. 36, pp. 717-731, 2020, doi:10.1007/s00371-019-01651-4.

[16] E. Iqbal *et. al.,* "Saliency-driven active contour model for image segmentation," *IEEE Access*, vol. 8, pp. 208978-208991, 2020, doi: 10.1109/ACCESS.2020.3038945.

[17] N. F. Idris, M. A. Ismail, M. S. Mohamad, S. Kasim, Z. Zakaria, and T. Sutikno, "Breast cancer disease classification using fuzzy ID3 algorithm based on association function," *IAES Int. J. Artif. Intell. (IJ-AI)*, vol. 11, no. 2, pp. 448, Jun. 2022, doi: 10.11591/ijai.v11.i2.pp448-461.

[18] H. M. Ahmed and M. Y. Kashmola, "A proposed architecture for convolutional neural networks to detect skin cancers," *IAES Int. J. Artif. Intell. (IJ-AI)*, vol. 11, no. 2, pp. 485, Jun. 2022, doi: 10.11591/ijai. v11.i2.

[19] A. H. T. Al-Ghrairi, A. A. Mohammed, and E. Z. Sameen, "Face detection and recognition with 180 degree rotation based on principal component analysis algorithm," *IAES Int. J. Artif. Intell.*, vol. 11, no. 2, pp. 593–602, 2022, doi: 10.11591/ijai.v11.i2.pp593-602.

[20] M. Z. Ansari, T. Ahmad, M. M. S. Beg, and N. Bari, "Language lexicons for Hindi-English multilingual text processing," *IAES Int. J. Artif. Intell. (IJ-AI)*, vol. 11, no. 2, pp. 641, 2022, doi: 10.11591/ijai. v11.i2. pp641-648.

[21] A. K. Jumaat and K. Chen, K, "A reformulated convex and selective variational image segmentation model and its fast multilevel algorithm," *Numerical Mathematics Theory Methods and Applications*, vol. 12, no. 2,pp. 403–437, 2019, doi: 0.4208/nmtma.OA-2017-0143.

[22] N. A. S. Mohd Ghani, A. K. Jumaat and R. Mahmud, "Boundary extraction of abnormality region in breast mammography image using active contours," *ESTEEM Academic J.*, vol. 18, pp.115-127, 2022.

[23] T. C. Saibin and A. K. Jumaat, "Variational selective segmentation model for intensity inhomogeneous image," *Indonesian J. Elect. Eng. and Comput. Sci.*, vol. 29, pp. 277-285, 2023, doi:1 10.11591/ijeecs.v29.i1.

[24] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso and J. S. Cardoso, "Toward a Full-field Digital Mammographic Database. Academic Radiology," Sep 2011, Distributed by INbreast. vol 19, no. 2, pp. 236–248. https://doi.org/10.1016/j.acra.2011.09.014.

[25] S. Candemir, S. Jaeger, K. Palaniappan, J.P. Musco, R. K. Singh, Z. Xue, A. Karargyris, S. Antani, G. Thoma and C. J. McDonald, "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," *IEEE Trans Med Imaging*, vol. 33, no. 2, pp.

577-90, 2014, doi: 10.1109/TMI.2013.2290491. PMID: 24239990.

[26] A. K. Jumaat and K. Chen, "Three-dimensional convex and selective variational image segmentation model," *Malaysian Journal of Mathematical Sciences*, vol. 14, no. 3, pp. 437–450, 2020, https://einspem.upm.edu.my/journal/fullpaper/vol14no3/7.%20Abdul%20Kadir%20Jumaat.pdf.

[27] A. K. Jumaat and K. Chen, "An optimization-based multilevel algorithm for variational image segmentation models," *Electronic Transactions on Numerical Analysis*, vol. 46, pp. 474–504, 2017, https://etna.ricam.oeaw.ac.at/vol.46.2017/pp474-504.dir/pp474-504.pdf.

[28] P. T. Krishnan, P. Balasubramanian, V. Jeyakumar, S. Mahadevan and A. Noel Joseph Raj, "Intensity matching through saliency maps for thermal and visible image registration for face detection applications," *Vis Comput.,* pp. 1–14, 2022, doi: 10.1007/s00371-022-02605-z.

[29] V. Lukin, E. Bataeva and S. Abramov, "Saliency map in image visual quality assessment and processing," *Radioelectronic and Computer Systems*, vol. 1, pp. 112–121, 2023, doi: 10.32620/reks.2023.1.09.

Biography of all authors

Picture	Biography	Authorship contribution
	Nadiah Mohamed obtained the B.Sc. degrees in Mathematics from the Universiti Teknologi MARA, (UiTM) Shah Alam, Malaysia and M.Sc. in Mathematics from Universiti Kebangsaan Malaysia. Presently, she holds the position of senior lecturer at the School of Mathematical Sciences in UiTM Negeri Sembilan Kampus Kuala Pilah, Malaysia. Her research interests include operational research, image/signal processing, and medical image analysis. Her email address is nadiah@uitm.edu.my.	Design the research work, data collection, data analysis and interpretation, drafting article.
	Abdul Kadir Jumaat obtained the B.Sc. and M.Sc. degrees in Mathematics from the UiTM Shah Alam, Malaysia and the University of Liverpool, in United Kingdom, awarded him a Ph.D. degree particularly in Applied Mathematics. Presently, he holds the position of senior lecturer at the School of Mathematical Sciences and a research fellow at the Institute for Big Data Analytics and Artificial Intelligence (IBDAAI) in UiTM Shah Alam, Malaysia. His research interests include image/signal processing, artificial intelligence and computer vision. His email address is abdulkadir@tmsk.uitm.edu.my	Design and supervising the research work and camera ready.
	Rozi Mahmud obtained the medical degree in University of Malaya. She subsequently went on to pursue her Masters in Radiology and obtained her MMed Rad from Universiti Kebangsaan Malaysia. She received a formal training at the Division of Neuroradiology and MRI at the Children's Hospital of Philadelphia and a formal training for mammogram and breast biopsy from L'Istituto Nazionale di Fisica Nucleare INFN, Sezione in Torino Italy. Presently, she holds the position of Professor and Director, Centre for Diagnostic Nuclear Imaging, Universiti Putra Malaysia. Her research interests include Radiology / Neuroradiology/Breast Disease/AI in medical Imaging. Her email address is rozi@upm.edu.my	Supervising the research work

Complex Network Analysis of the NASDAQ Stock Market during the Initial Phase of COVID-19

Chin Jia Hou
Department of Mathematical and Actuarial Sciences, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, 43000 Kajang, Selangor, Malaysia
jhchin2013@gmail.com

Nor Aziyatul Izni
Centre of Foundation Studies, Universiti Teknologi MARA, Cawangan Selangor, Kampus Dengkil, 43800 Dengkil, Selangor, Malaysia
naizni@uitm.edu.my

ABSTRACT

The United States (U.S.) plays an important role in the global economy, and the COVID-19 pandemic significantly affected the U.S. stock market. Over the past two decades, numerous studies have incorporated complex network analysis to analyze the stock market. However, there is a lack of study focused on identifying anomalies in the complex network structure of the U.S. stock market that could indicate impending financial crises. The main objective of this research is to implement complex network analysis in examining the changes in the network structures and centralities of the NASDAQ stock networks leading up to and during the initial phase of the COVID-19 pandemic. The opening prices of the stocks under the NASDAQ index in the last two quarters of 2019 and the first quarter of 2020 were collected from Yahoo Finance. The collected data was parsed into edges lists which were then used to construct multiple stock networks. The structures of the stock networks were analyzed using topological metrics such as network density, average clustering coefficient, average path length, network centralizations, and modularity of community structure. The centrality scores of the stocks in the networks were calculated and they were ranked according to the scores. The results show abnormal values in the number of edges, network density, betweenness centralization, and modularity of the community structure during the initial phase of the COVID-19 pandemic. However, no significant anomalies are observed in the average clustering coefficient, average path length, degree centralization, and closeness centralization. Meanwhile, degree centrality proves effective in identifying influential stocks, while closeness and betweenness centralities are found to be less suitable for this particular purpose in the networks used in this study. This paper provides insights into the changes within the stock market at both micro and macro levels around the financial crisis, where the anomalies serve as indicators of an impending financial crisis.

Corresponding Author:
Name: Chin Jia Hou
Affiliation: Department of Mathematical and Actuarial Sciences, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, 43000 Kajang, Selangor, Malaysia
Email: chinjh2013@gmail.com

1. Introduction

Complex network analysis or network science is an emerging research field dedicated to exploring the properties and features inherent in complex networks. In general, complex networks consist of nodes denoting distinct entities, and edges describing the relationship amid these entities. The nature of the nodes and their respective relationships varies across individual networks. For instance, in a communication network, nodes are defined as Twitter users, and edges are used to link the users mentioned in a tweet [1]. In a container transportation network, nodes represent ports

while edges denote the movements of vessels between ports [2]. Complex network analysis proved to be a feasible approach to providing insightful findings. For example, the optimization of electric power systems using complex network analysis can provide boosts in grid resilience in the event of a blackout [3]. On the other hand, the application of complex network analysis in managing water supply systems can reduce construction costs, while improving the service quality [4].

The United States (U.S.) holds an important position in the global economy. Thus, stock traders always need to closely analyze and respond to the U.S. stock market. The outbreak of COVID-19 had a profound effect on the U.S. stock market, sending momentous shock waves on indices such as the NASDAQ index, which is one of the primary stock indices in the U.S. Substantially, the interconnectedness of the U.S. stock market with the global financial system emphasizes the far-reaching effect of events such as the COVID-19 pandemic on stock markets.

The stock market comprises a lot of frequently interacting stocks. Hence, it can be considered as a complex system. Numerous research works were dedicated to understanding the stock market through complex network analysis. By examining patterns in network properties, the underlying structure of stock markets can be revealed. Furthermore, influential stocks can be identified using network centrality for further analysis. Despite the feasibility of complex network analysis in studying stock markets, there is a lack of studies utilizing this approach to investigate the evolution of stock behavior under the NASDAQ index during financial crises. Studies indicate that the stock market behaves differently during financial crises [5], [6]. By understanding the changes within the stock market at both micro and macro levels, anomalies could be identified which then serve as indicators of an impending crisis.

In this study, we propose the implementation of complex network analysis to examine the structure of the NASDAQ stock market leading up to and during the initial phase of the COVID-19 pandemic. This involves the analysis of the changes in network structure via the network metrics, including network density, average clustering coefficient, average path length, network centralization, and modularity of community structure. Furthermore, we investigated the variation in the influence of individual stocks by assessing their centralities in the networks.

2. Literature Review

The crucial component of complex network analysis lies in the construction of the network of interest by defining the relationship between entities in a complex system. In stock networks, the entities are stocks and their correlation in terms of their prices or volumes will define the connections between them. One of the most used models to capture the correlations among stocks is Pearson's correlation with threshold [7], [8].

Node centralities in stock networks can highlight the influential stocks in the networks. Wang et. al. [9] utilized PageRank to identify influential energy stocks under 30 China financial institutions. In a study on the Shanghai and Shenzhen stock markets, it was found that stocks with smaller market capitalization have more central positions in networks [10]. Similar phenomena can be observed in the Iranian stock market, where stocks with higher centralities have higher market capitalization and price fluctuation, as well as larger transaction volume [11]. Furthermore, Wang et. al. [12] investigated the correlation between centralities and financial indicators to find out the centralities that contribute the most to identifying influential stocks.

Aside from centralities, by observing the changes in the network topology and structure, the behavior of the stock market could be unveiled [13], [14]. Huang et. al. [15] proposed a new indicator to detect subprime crisis, which has better performance than classical indicators such as degree and clustering coefficient. For financial network analysis, the changes in the distribution and average of the centralities in dynamic networks (i.e. networks that evolve through time) reveal the evolution of behaviors of the stock market during the financial crisis and normal periods [16]. Community structure within a network provides insights into the presence of stock clusters which can be focused on to avert potential damages to the entire stock market [17]. By using a modularity-based community detection method, Chen et. al. [10] categorized A-Share stocks by industries and measured the intra- and inter-industry connectivity of the stocks. Numerous research works suggested that stock complex networks exhibit scale-free properties, where the degree distribution follows a power law distribution [18] - [20]. Scale-free networks are complex networks that show a preferential attachment process. The main characteristic of this kind of network is the existence of "hub" nodes with exceptionally high degrees than the other nodes.

Complex network analysis has been applied to the study of the U.S. stock market. Aslam et. al. [21] investigated the centralities of global stock market indices in the pre- and post-COVID-19

periods, and the results suggested that the U.S. index (DOW 30) did not lead global stock markets before or during the COVID-19 period. In a study of global stock market co-movement during the COVID-19, Huang et. al. [22] showed that the centrality rankings of the U.S. stock price index, based on degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality, changed only marginally from 2020 to 2022. Furthermore, the U.S. maintained a stable position in the global stock market network throughout this period.

Korkusuz et al. [23] analyzed volatility spillovers during financial crises using network centralities, showing that the U.S. was a major source of financial volatility during the Global Financial Crisis. Furthermore, this volatility remained significant during the COVID-19 Crisis, increasing the overall transitivity of the volatility network. Li and Pi [24] investigated the clustering coefficients and community structure within the complex networks of global stock indices. They found that the U.S. stock market exhibited regional clustering, particularly during financial crises. The correlation among global stock indices, including U.S. indices, increased during such periods.

3. Methodology

3.1 Data Collection and Network Construction

Stock data of the stocks under the NASDAQ index were retrieved from the Yahoo Finance website using the "quantmod" package in R. Aside from the symbols and security names of the stocks, their daily opening prices from 1 July 2019 to 31 March 2020 were retrieved. In total, data for 3063 stocks were collected. The cross-correlations between the stocks were calculated. Let $x_i(t)$ and $x_j(t)$ be the daily opening prices of stock i and j, respectively, over the period $t = 0$ to $t = n - 1$. The cross-correlation between the stocks with no time shift is defined as [25] depicted in Equation (1).

$$c_{ij} = \frac{\Sigma_t[(x_i(t) - \bar{x}_i)(x_j(t) - \bar{x}_j)]}{\sqrt{\Sigma_t(x_i(t) - \bar{x}_i)^2}\sqrt{\Sigma_t(x_j(t) - \bar{x}_j)^2}}$$

(1)

where $\bar{x}_i$ and $\bar{x}_j$ are the means of the time series, and c_{ij} ranges from 0 to 1.

Subsequently, stock networks were constructed by using cross-correlations, where the nodes represent stocks that are connected by edges if c_{ij}, which serves as the edge weight, is a positive value. Due to the excessively high number of edges in the networks constructed without restriction in c_{ij}, a winner-take-all approach is employed to reduce the networks. Specifically, three threshold values (T1 = 0.7; T2 = 0.8; T3 = 0.9) were chosen. Edges with weights (c_{ij}) less than the threshold values were removed from the networks. Furthermore, the stock dataset is divided into three periods, namely 2019_Q3 (2851 out of 3063 stocks) for 1 Jul 2019 to 30 Sept 2019 (quarter 3 of year 2019), 2019_Q4 (2889 out of 3063 stocks) for 1 Oct 2019 to 31 Dec 2019 (quarter 4 of year 2019), and 2020_Q1 (2923 out of 3063 stocks) for 1 Jan 2020 to 31 Mar 2020 (quarter 1 of year 2020). Hence, nine stock networks that depict the cross-correlations of the stocks at various periods and cross-correlations threshold values were constructed: 2019_Q3_T1, 2019_Q3_T2, 2019_Q3_T3, 2019_Q4_T1, 2019_Q4_T2, 2019_Q4_T3, 2020_Q1_T1, 2020_Q1_T2, 2020_Q1_T3. Since there is no direction when defining the cross-correlations of the stocks, the stock networks are undirected and weighted. Gephi was used to visualize the stock networks.

3.2 Network Theory

The density of a network quantifies the proportion of the actual number of edges relative to the total possible number of edges in a network. For an undirected network, network density is defined as Equation (2).

$$ND = \frac{2|E|}{N(N - 1)}$$

(2)

where $|E|$ is the number of edges while N is the number of nodes in the network. The density of a network ranges between 0 and 1, with higher values indicating stronger interconnectedness between the nodes in the network.

The clustering coefficient measures the probability of connection between the neighboring nodes of a node. The local clustering coefficient of node i in an undirected and weighted network is defined as [26] depicted in Equation (3).

$$CC_i^w = \frac{1}{s_i(k_i - 1)} \sum_{j,h} \frac{w_{ij} + w_{ih}}{2} a_{ij} a_{ih} a_{jh}$$
(3)

where s_i is the strength of node i, a_{ij} is an element of the adjacency matrix (row i column j), k_i is the degree of node i, and w_{ij} is the weight of the edge connecting node i and j. The average clustering coefficient of a network is the mean of the local clustering coefficients. It ranges from 0 to 1, with higher values indicating higher tendency of nodes in the network to cluster together.

The average path length denotes the average number of steps that are required to move from one node to another in a network. It is calculated by averaging the lengths of the shortest paths between all pairs of nodes in the network [27]. The shortest path lengths in weighted networks can be calculated using Dijkstra's algorithm. The average path length offers a glimpse of the connectivity of the network.

In general, the centralities of nodes quantify the importance of the nodes within a network. The degree centrality, closeness centrality, and betweenness centrality [28] are the most used centrality measures in complex network analysis. The centralities of node i in an undirected and weighted network are defined as:

$$\text{Degree centrality: } C_D(i) = \sum_j a_{ij} = k_i$$
(4)

$$\text{Closeness centrality: } C_C(i) = \frac{N-1}{\sum_{j=1}^{N} d(i,j)}$$
(5)

$$\text{Betweenness centrality: } C_B(i) = \sum_{i \neq j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$$
(6)

where $d(i,j)$ is the shortest path length between nodes i and j, σ_{jk} is the total number of shortest paths between nodes j and k, and $\sigma_{jk}(i)$ is the number of shortest paths between nodes j and k that go through node i. Closeness centrality and betweenness centrality can be normalized to range between 0 and 1. Each centrality is interpreted differently. A node with a high degree centrality is important as it has a lot of connections. On the other hand, a node with high closeness centrality can swiftly interact with all other nodes in a network. Meanwhile, a node with high betweenness centrality plays a crucial role as a bridge that connects various components within the network.

Centralization of a network is a network-level metric derived from the centrality scores of individual nodes, enabling the comparison of different networks. Essentially, when network centralization is high, there is a greater likelihood that a single node holds a central position within the network. Let X_i be the degree, closeness, or betweenness centrality, the centralization is then defined as Equation (7).

$$X_g = \frac{\sum_{i=1}^{N}(X^* - X_i)}{\max \sum_{i=1}^{N}(X^* - X_i)}$$
(7)

where $X^* = \max(X_i)$. D_g, C_g and B_g are used to denote degree, closeness, and betweenness centralities, respectively.

The Leiden algorithm was implemented in the detection of communities in the networks [29]. Modularity (Q) was utilized to gauge the quality of the detected communities [30]. The value of Q ranges from 0 to 1, with high Q values signifying a strong community structure within a network.

4. Results and Discussion

4.1 Network Topological Metrics of the NASDAQ Stock Networks

Table 1 depicts the changes in network metrics across different cross-correlation threshold values over consecutive periods, while the visualization of the 2019_Q3_T3, 2019_Q4_T3, and 2020_Q1_T3 stock networks are depicted in Figures 1 to 3.

Given that the threshold values determine the presence of edges between stocks in the networks, it is natural that the number of nodes and edges diminishes with an increase in the threshold value. Notably, changes in the number of nodes across the threshold values are marginal, except for a 26% reduction in the number of nodes during 2019 Q3 at the threshold value of 0.9. An interesting observation is the sudden surge in the number of edges during 2020 Q1, marking the initial phase of the COVID-19 pandemic. This increase is consistent across all threshold values. Specifically, the number of edges during 2020 Q1 surpasses that of 2019 Q4 by 2.8 times, 4.3 times, and 8.9 times at the corresponding threshold values. The onset of the COVID-19 pandemic brought about widespread market shock, inducing analogous responses among investors dealing with economic uncertainty. This collective reaction is reflected in the intensified connectivity among stocks during this period.

The network densities observed in 2019 Q3 and 2019 Q4 are below 0.3, indicating a low level of interconnectivity among stocks during those periods. This suggests a lack of noticeable overall co-movement among the stocks. However, the advent of the COVID-19 pandemic significantly raises the network density in 2020 Q1, particularly at the threshold value of 0.9. The notable increase aligns with the substantial rise in the number of edges during this period. It can be observed from Figures 1 to 3 that the networks are getting denser with the increase in the number of edges connecting the nodes in the networks.

Table 1. The statistics and network metrics of the stock networks at various periods and cross-correlation threshold values.

Networks	N	$|E|$	ND	ACC	APL	D_g	C_g	B_g	Q
Threshold Value = 0.7									
2019_Q3_T1	2745	476353	0.126	0.687	1.801	0.338	0.349	0.163	0.367
2019_Q4_T1	2856	868357	0.213	0.768	1.656	0.648	0.664	0.083	0.073
2020_Q1_T1	2894	3336955	0.797	0.938	1.043	0.137	0.177	0.032	0.004
Threshold Value = 0.8									
2019_Q3_T2	2549	224516	0.069	0.649	2.353	0.306	0.345	0.298	0.355
2019_Q4_T2	2848	533450	0.132	0.766	1.886	0.731	0.699	0.099	0.043
2020_Q1_T2	2863	2821094	0.689	0.896	1.194	0.215	0.263	0.034	0.007
Threshold Value = 0.9									
2019_Q3_T3	1883	41482	0.023	0.572	3.367	0.241	0.319	0.528	0.310
2019_Q4_T3	2834	166507	0.041	0.803	2.157	0.826	0.707	0.125	0.006
2020_Q1_T3	2756	1653329	0.436	0.810	1.562	0.331	0.340	0.051	0.010

Notes: N and $|E|$ are the numbers of nodes and edges in the network, while ND, ACC, APL, D_g, C_g and B_g denote the density, average clustering coefficient, average path length, degree, closeness, and betweenness centralizations of the networks, respectively. Q is the modularity of the detected communities.
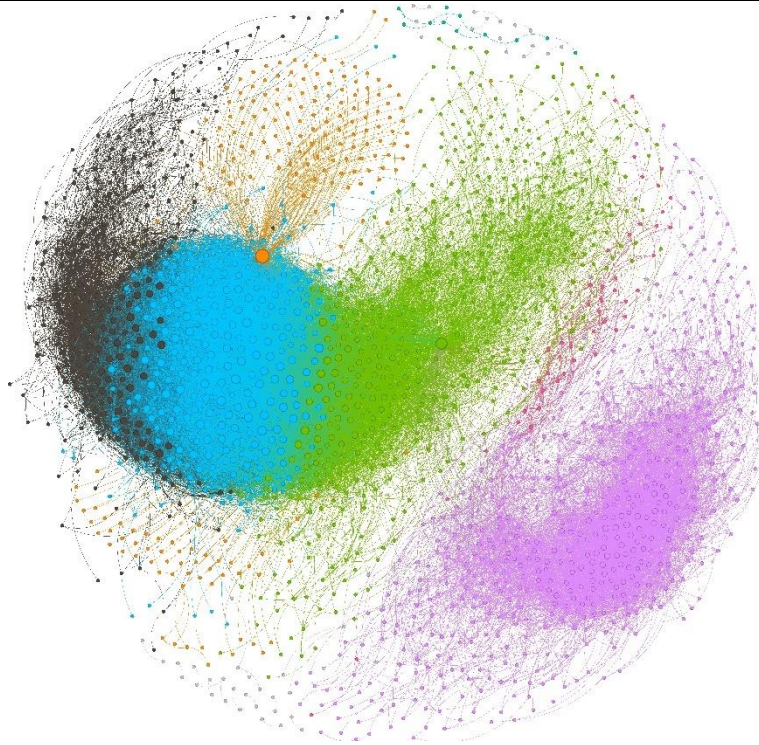
Figure 1. Visualization of the NASDAQ stock network during 2019 quarter 3 at the threshold value of 0.9 (2019_Q3_T3). The size of the nodes denotes the degree centrality score of the nodes. The nodes are colored by communities.
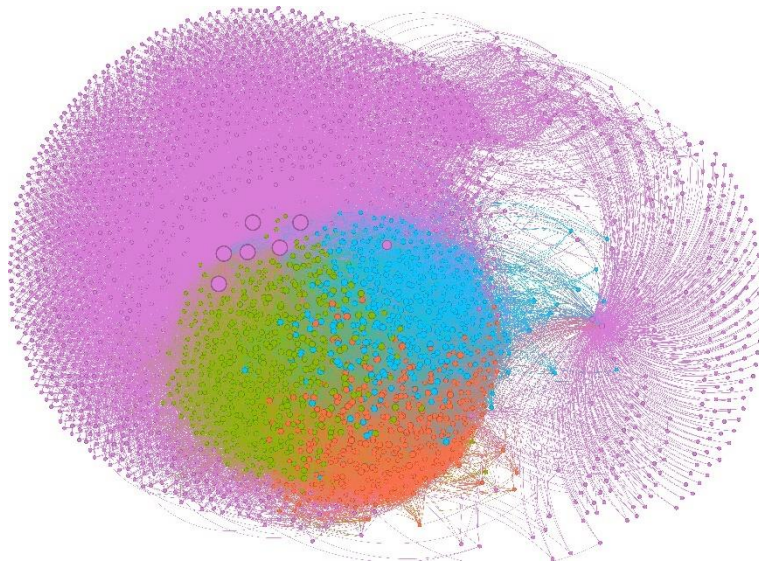
Figure 2. Visualization of the NASDAQ stock network during 2019 quarter 4 at the threshold value of 0.9 (2019_Q4_T3). The size of the nodes denotes the degree centrality score of the nodes. The nodes are colored by communities.
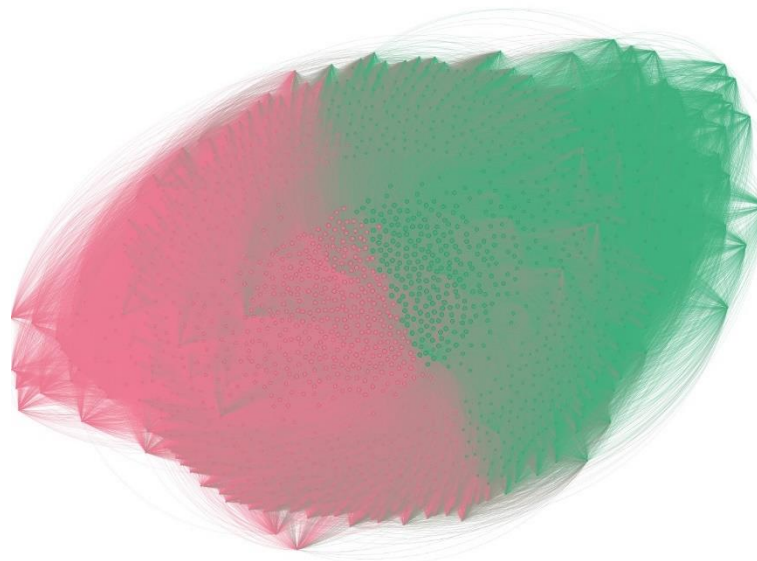
Figure 3. Visualization of the NASDAQ stock network during 2020 quarter 1 at the threshold value of 0.9 (2020_Q1_T3). The size of the nodes denotes the degree centrality score of the nodes. The nodes are colored by communities.

Although the average clustering coefficient is the highest in 2020 Q1, the increment is gradual over the periods. Specifically, at the threshold value of 0.9, the average clustering coefficient increases merely by 0.8% from 2019 Q3 to 2020 Q4. Given the diverse sectors encompassed by NASDAQ, including energy, financials, healthcare, technology, and telecommunications, it is foreseeable that stocks within the same sector form tightly knit clusters, showing strong interconnectedness within sectors. However, despite the tendency of stocks to cluster together due to the high average clustering coefficient, the low modularity scores (Q) of the identified community structure in 2020 Q1 indicate a diminished community structure during the economic turbulence caused by COVID-19. As displayed in Figure 1, the communities in the 2019_Q3_T3 network are distinguishable, with one cluster of stocks (purple) separated from the other clusters. On the other hand, all the stocks are tightly clustered together in the 2020_Q1_T3 network, with no clear boundary between the two detected communities (pink and green). Hence, in 2020 Q1, the co-movement of stocks is observed on an overall scale, rather than within sectors. Nonetheless, it is interesting to note that the community structure in 2019 Q4 is as weak as in 2020 Q1. Further investigation into the network structure of the stocks during this period is needed to gain insights into this observation.

Examining the average path lengths of stock markets provides insights into the speed at which shocks spread across networks. In general, NASDAQ stock networks exhibit very small average path lengths, with the largest value reported at 3.367 for the threshold value of 0.9 in 2019 Q3. Coupled with the relatively large average clustering coefficients in the corresponding networks, it indicates that the NASDAQ stock networks demonstrate a small-world property [31]. Stock networks with such property are more sensitive to systemic risk and demonstrate distinct co-movements during financial crises.

The exceptionally low betweenness centralization values across all threshold values in the stock networks emphasize the influence of COVID-19 on the stock market structure. In 2020 Q1, the market was decentralized in terms of betweenness centrality, with most stocks possessing similar betweenness centrality scores. This suggests that all sectors under NASDAQ were affected simultaneously during the initial phase of COVID-19, instead of spreading from sector to sector. In contrast, the changes in degree and closeness centralizations during 2020 Q1 are not as pronounced as those observed in betweenness centralization. The relatively low network centralization values across various threshold values and periods indicate that NASDAQ has a decentralized market structure. However, in 2019 Q4, degree, and closeness centralizations are relatively high, indicating the existence of a few stocks with a high number of connections that could rapidly influence a large portion of NASDAQ stocks. This can be observed in Figure 2, where there are 5 purple nodes which are significantly larger than the other nodes in the network.

4.2 Identification of Influential Stocks in the NASDAQ Stock Networks

The degree centrality of stocks within the stock networks serves as a measure of their influence regarding the co-movement of stocks in the stock market. Table 2 shows the top 5 NASDAQ stocks ranked by degree centrality.

Table 2. The top 5 NASDAQ stocks ranked by degree centrality at various periods and cross-correlation threshold values. The stocks are represented by their symbols.

Rank	2019_Q3_T1	2019_Q3_T2	2019_Q3_T3
1	TFII (1275) (Logistic)	TFII (956) (Logistic)	DTP (498) (Utilities)
2	DTP (971) (Utilities)	DTP (835) (Utilities)	TFII (381) (Logistic)
3	LNC (862) (Finance)	CIT (617) (Finance)	PUK (278) (Finance)
4	FLS (854) (Industrials)	MET (606) (Finance)	NGVT (275) (Industrials)
5	MET (849) (Finance)	NGVT (603) (Industrials)	CIT (269) (Finance)
Rank	**2019_Q4_T1**	**2019_Q4_T2**	**2019_Q4_T3**
1	BFYT (2457) (Finance)	BFYT (2457) (Finance)	BFYT (2457) (Finance)
2	BPYU (2457) (Real Estate)	BPYU (2457) (Real Estate)	BPYU (2457) (Real Estate)
3	BPYUP (2457) (Real Estate)	BPYUP (2457) (Real Estate)	BPYUP (2457) (Real Estate)
4	PIPR (2457) (Finance)	PIPR (2457) (Finance)	PIPR (2457) (Finance)
5	TFII (2457) (Logistic)	TFII (2457) (Logistic)	TFII (2457) (Logistic)
Rank	**2020_Q1_T1**	**2020_Q1_T2**	**2020_Q1_T3**
1	WTRG (2703) (Utilities)	TFII (2585) (Logistic)	TFII (2113) (Logistic)
2	TFII (2697) (Logistic)	QQQX (2536) (Finance)	DIAX (2055) (Finance)
3	MGU (2666) (Finance)	CAPE (2532) (Finance)	GDV (2053) (Finance)
4	IPG (2664) (Consumer Discretionary)	GAM (2532) (Finance)	AIR (2034) (Industrials)
5	MORN (2664) (Finance)	PIPR (2530) (Finance)	DOV (2034) (Industrials)

Notes: The first bracket in each cell denotes the degree centrality scores, while the second bracket denotes the sector.

As shown in Table 2, TFI International Inc. (TFII), a logistics company, emerges as an influential stock across all threshold values and periods. This observation is reasonable considering the important role logistics plays in the supply chain, affecting various sectors and industries. Furthermore, a general trend is observed where stocks within the finance sector are influential during the periods investigated in this study, especially during the early phase of the COVID-19 pandemic (2020 Q1).

Upon closer examination of the degree centrality scores for all stocks in 2019 Q4, it can be noticed that the top 5 influential stocks, namely Benefytt Technologies, Inc. (BFYT), Brookfield Property REIT Inc. - Class A (BPYU), Brookfield Property REIT Inc. - 6.375% Series A (BPYUP), Piper Sandler Companies (PIPR), and TFI International Inc. (TFII), possess higher degree centrality scores than the remaining stocks in the networks. This difference contributes to the relatively high degree centralization during 2019 Q4, as discussed in the previous subsection. Notably, the top 5 influential stocks during 2019 Q4 remained the same across all threshold values.

On the other hand, it can be observed from Table 3 that the highest closeness and betweenness centrality scores in all the stock networks are remarkably low (falling below 0.1), except the betweenness centrality of TFI International Inc. (TFII) in the 2019_Q3_T2 and 2019_Q3_T3 networks, Immunovant, Inc. (IMVTU) in the 2019_Q4_T2 network, and Benefytt Technologies, Inc. (BFYT) in the 2019_Q4_T3 network. This observation implies that even though stocks can be ranked based on closeness and betweenness centralities, they are not considered influential, given the low centrality scores during the periods examined in this study.

Table 3. The stocks with the highest closeness and betweenness centralities at various periods and cross-correlation threshold values.

Networks	Highest Closeness Centrality	Highest Betweenness Centrality
2019_Q3_T1	TFII (0.00029)	TFII (0.08998)
2019_Q3_T2	TFII (0.00025)	TFII (0.21752)
2019_Q3_T3	TFII (0.00026)	TFII (0.49398)
2019_Q4_T1	BFYT (0.00031)	IMVTU (0.07384)
2019_Q4_T2	BFYT (0.00031)	IMVTU (0.10027)
2019_Q4_T3	BFYT (0.00029)	BFYT (0.10977)
2020_Q1_T1	IPHI (0.00038)	RFM (0.02560)
2020_Q1_T2	BSTZ (0.00034)	RFM (0.02590)
2020_Q1_T3	TFII (0.00030)	RFM (0.04658)

Notes: The stocks are represented by their symbols and the values in the brackets represent the corresponding degree centrality scores

5. Conclusion

In this paper, the structure of the NASDAQ stock market and the co-movement behaviors of NASDAQ stocks around COVID-19 were analyzed through the lens of complex network analysis. The findings reveal abnormal values in the number of edges, network density, betweenness centralization, and modularity of the community structure during the early phase of COVID-19. While the average clustering coefficient, average path length, degree centralization, and closeness centralization do not exhibit distinctive anomalies during this period, they offer valuable insights into the intrinsic structure of the NASDAQ stock market. The application of various centralities on the NASDAQ stock networks indicates that degree centrality can effectively identify influential stocks, whereas closeness and betweenness centralities are less suitable for this purpose.

For future research, it would be interesting to observe the structural changes in the NASDAQ stock market across different phases of COVID-19. Furthermore, extending this study to other stock markets such as S&P500 and Dow Jones could provide a more comprehensive understanding of the U.S. stock market during financial crises.

Conflict of Interest
The authors declare no conflict of interest in the subject matter or materials discussed in this manuscript.

References
[1]   P. Pascual-Ferrá, N. Alperstein, and D. J. Barnett, "Social network analysis of COVID-19 public discourse on Twitter: Implications for risk communication," *Disaster Med Public Health Prep*, vol. 16, no. 2, pp. 561–569, 2022.
[2]   D. Guerrero, L. Letrouit, and C. Pais-Montes, "The container transport system during Covid-19: An analysis through the prism of complex networks," *Transp Policy (Oxf)*, vol. 115, pp. 113–125, 2022.
[3]   M. Saleh, Y. Esa, and A. Mohamed, "Applications of complex network analysis in electric power systems," *Energies (Basel)*, vol. 11, no. 6, p. 1381, 2018.
[4]   A. Yazdani and P. Jeffrey, "Complex network analysis of water distribution systems," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 21, no. 1, 2011.
[5]   C. O. Cepoi, "Asymmetric dependence between stock market returns and news during COVID-19 financial turmoil," *Financ Res Lett*, vol. 36, Oct. 2020, doi: 10.1016/j.frl.2020.101658.

[6]   J. W. Goodell and T. L. D. Huynh, "Did Congress trade ahead? Considering the reaction of US industries to COVID-19," in *Finance Research Letters*, Elsevier Ltd, Oct. 2020. doi: 10.1016/j.frl.2020.101578.

[7]   F. N. A. Mahamood, H. Bahaludin, M. H. Abdullah, and others, "A network analysis of Shariah-compliant stocks across global financial crisis: A case of Malaysia," *Mod Appl Sci*, vol. 13, no. 7, pp. 80–93, 2019.

[8]   B. A. Memon, H. Yao, F. Aslam, and R. Tahir, "Network analysis of Pakistan stock market during the turbulence of economic crisis," *Business, Management and Economics Engineering*, vol. 17, no. 2, pp. 269–285, 2019.

[9]   Z. Wang, X. Gao, H. An, R. Tang, and Q. Sun, "Identifying influential energy stocks based on spillover network," *International Review of Financial Analysis*, vol. 68, p. 101277, 2020.

[10]   K. Chen, P. Luo, B. Sun, and H. Wang, "Which stocks are profitable? A network method to investigate the effects of network structure on stock returns," *Physica A: Statistical Mechanics and its Applications*, vol. 436, pp. 224–235, 2015.

[11]   H. E. Moghadam, T. Mohammadi, M. F. Kashani, and A. Shakeri, "Complex networks analysis in Iran stock market: The application of centrality," *Physica A: Statistical Mechanics and its Applications*, vol. 531, p. 121800, 2019.

[12]   H. Wang, Z. Du, J. M. Moore, H. Yang, and C. Gu, "Causal networks reveal the response of Chinese stocks to modern crises," *Inf Sci (N Y)*, vol. 609, pp. 1670–1693, 2022, doi: https://doi.org/10.1016/j.ins.2022.07.159.

[13]   Y. Zhou, Z. Chen, and Z. Liu, "Dynamic analysis and community recognition of stock price based on a complex network perspective," *Expert Syst Appl*, vol. 213, p. 118944, 2023, doi: https://doi.org/10.1016/j.eswa.2022.118944.

[14]   K. Zaheer, F. Aslam, Y. Tariq Mohmand, and P. Ferreira, "Temporal changes in global stock markets during COVID-19: an analysis of dynamic networks," *China Finance Review International*, 2022, doi: 10.1108/CFRI-07-2021-0137.

[15]   X. Y. Chuangxia Huang Shijie Liu and X. Yang, "Identification of crisis in the Chinese stock market based on complex network," *Appl Econ Lett*, vol. 30, no. 18, pp. 2536–2542, 2023, doi: 10.1080/13504851.2022.2099792.

[16]   Y. Lai and Y. Hu, "A study of systemic risk of global stock markets under COVID-19 based on complex financial networks," *Physica A: Statistical Mechanics and its Applications*, vol. 566, p. 125613, 2021, doi: https://doi.org/10.1016/j.physa.2020.125613.

[17]   S. Liu, M. Caporin, and S. Paterlini, "Dynamic network analysis of North American financial institutions," *Financ Res Lett*, vol. 42, p. 101921, 2021.

[18]   M. A. Djauhari and S. L. Gan, "Optimality problem of network topology in stocks market analysis," *Physica A: Statistical Mechanics and its Applications*, vol. 419, pp. 108–114, 2015.

[19]   T. You, P. Fiedor, and A. Hołda, "Network analysis of the Shanghai stock exchange based on partial mutual information," *Journal of Risk and Financial Management*, vol. 8, no. 2, pp. 266–284, 2015.

[20]   B. A. Memon, H. Yao, F. Aslam, and R. Tahir, "Network Analysis of Pakistan Stock Market during the Turbulence of Economic Crisis," *Business, Management and Education*, vol. 17, no. 2, pp. 269–285, Dec. 2019, doi: 10.3846/bme.2019.11394.

[21]   F. Aslam, Y. T. Mohmand, P. Ferreira, B. A. Memon, M. Khan, and M. Khan, "Network analysis of global stock markets at the beginning of the coronavirus disease (Covid-19) outbreak," *Borsa Istanbul Review*, vol. 20, pp. S49–S61, 2020.

[22]   W. Huang, H. Wang, Y. Wei, and J. Chevallier, "Complex network analysis of global stock market co-movement during the COVID-19 pandemic based on intraday open-high-low-close data," *Financial Innovation*, vol. 10, no. 1, Dec. 2024, doi: 10.1186/s40854-023-00548-5.

[23]   B. Korkusuz, D. G. McMillan, and D. Kambouroudis, "Complex network analysis of volatility spillovers between global financial indicators and G20 stock markets," *Empir Econ*, vol. 64, no. 4, pp. 1517–1537, Apr. 2023, doi: 10.1007/s00181-022-02290-w.

[24]   B. Li and D. Pi, "Analysis of global stock index data during crisis period via complex network approach," *PLoS One*, vol. 13, no. 7, Jul. 2018, doi: 10.1371/journal.pone.0200600.

[25]   J. Cohen, P. Cohen, S. G. West, and L. S. Aiken, *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge, 2013.

[26]   A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *Proceedings of the National Academy of Sciences*, vol. 101, no. 11, pp. 3747–3752, 2004.

[27]   D. B. West, *Introduction to graph theory*, vol. 2. Prentice Hall Upper Saddle River, 2001.

[28] L. C. Freeman and others, "Centrality in social networks: Conceptual clarification," *Social Network: Critical Concepts in Sociology. Londres: Routledge*, vol. 1, pp. 238–263, 2002.

[29] V. A. Traag, L. Waltman, and N. J. Van Eck, "From Louvain to Leiden: guaranteeing well-connected communities," *Sci Rep*, vol. 9, no. 1, p. 5233, 2019.

[30] M. E. J. Newman, "Detecting community structure in networks," *Eur Phys J B*, vol. 38, pp. 321–330, 2004.

[31] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

Biography of all authors

Picture	Biography	Authorship contribution
	Chin Jia Hou is an assistant professor in Universiti Tunku Abdul Rahman, Sungai Long Campus. He graduated in B.Sc. in the Computational and Industrial Mathematics from the Department of Mathematical Sciences, University of Malaya, Kuala Lumpur in 2008. He also obtained his M.Sc. in atomic collision theory and Ph.D. in complex network analysis from the same university in 2012 and 2018. His research interest is network science, particularly social networks and community structure of complex networks.	Design the research work, data analysis and interpretation, drafting article.
	Nor Aziyatul Izni is a senior lecturer at the Centre of Foundation Studies, Universiti Teknologi MARA, Selangor Branch, Dengkil Campus. She is awarded a Professional Technologist by the Malaysia Board of Technologists since November 2021. She graduated in Bachelor of Applied Science (Honours) in Mathematical Modelling in 2012, and a Master of Science (Mathematics) in 2013 from Universiti Sains Malaysia, Malaysia. She then completed her Ph.D. in 2019 from Universiti Teknologi Malaysia, Kuala Lumpur. Her research interests are data analytics, applied mathematics and statistics, signal processing, fusion algorithm, artificial intelligence as well as prediction and forecasting methods.	Data processing, data analysis and interpretation, drafting article.

Predicting Default and Non-Default Firms using Discriminant Analysis: Adaptation of KMV-Merton's Default Probabilities and Financial Ratios

Nur Ain Al-Hameefatul Jamaliyatul
Universiti Teknologi MARA, Cawangan Johor, Kampus Segamat, Johor, Malaysia.
ainnurain06@gmail.com

Nurul Afiqah Zainuddin
Unit 25-01, Level 25, Menara Felda, 11 Persiaran KLCC, 50450 Kuala Lumpur, Malaysia.
afiqahzainuddin99@gmail.com

Izza Suraya Zulhazmi
Wisma United Malacca, No.61 Level 4 Jalan Melaka Raya 8, Taman Melaka Raya, 75009 Melaka, Malaysia.
izzasuraya99@gmail.com

Norliza Muhamad Yusof
Kolej Pengajian Perkomputeran, Informatik dan Matematik, Universiti Teknologi MARA, Cawangan Negeri Sembilan, Kampus Seremban, Negeri Sembilan, Malaysia.
norliza3111@uitm.edu.my

Muhamad Luqman Sapini
Kolej Pengajian Perkomputeran, Informatik dan Matematik, Universiti Teknologi MARA, Cawangan Negeri Sembilan, Kampus Seremban, Negeri Sembilan, Malaysia.
luqman0211@uitm.edu.my

Article Info

ABSTRACT

The KMV-Merton model provides conceptual determinants for predicting firms' default risk, but its accuracy was tested long ago, and it contains insufficient statistics for default prediction. Therefore, previous literature adapted the KMV-Merton model into a statistical model involving financial ratios to improve its predictive capabilities. Discriminant Analysis (DA) is a widely used statistical model for predicting financial distress. The objectives of this study are to identify financial ratios significant to KMV-Merton's default probabilities using DA, to predict default and non-default firms using the DA model obtained, and to compare the performance of the KMV-Merton and DA models in predicting default risk. The study uses 11 years of data from Malaysian publicly listed firms, applying the KMV-Merton model and stepwise DA in SPSS. DA identifies the significance of selected financial ratios to firm default, with KMV-Merton's default probabilities as the dependent variable, forming a discrimination function to predict default and non-default firms. Credit ratings and Type 1 and Type II errors are used to compare model performance. The DA using SPSS reveals a discriminant function with net profit margin and return on assets significantly related to KMV-Merton's default probabilities. The DA model is more biased in predicting non-default firms due to the need for more information on default firms, yet it slightly outperforms the KMV-Merton model. This study offers guidance on adapting KMV-Merton's default probability estimates with financial ratios in the DA model and highlights the significant financial ratios related to KMV-Merton's default probabilities.

Corresponding Author:
Norliza Muhamad Yusof
Kolej Pengajian Perkomputeran, Informatik dan Matematik, Universiti Teknologi MARA Cawangan Negeri Sembilan Kampus Seremban, Negeri Sembilan, Malaysia.
Email: norliza3111@uitm.edu.my

PENERBIT PRESS
UNIVERSITI TEKNOLOGI MARA

1. Introduction

Default risk prediction is crucial as it helps to identify the financial condition of the firms [1] plus it assists the investors in acknowledging the firm's performance. Over the last decade, a variety of mathematical modelling methods have been initiated to estimate firms' default risk. Many categories of models can be separated based on the approach they take. The most well-known structural model of default risk, named KMV-Merton model, has been commonly used to forecast a firm's default. The KMV-Merton model has been used since the early period of credit risk modelling.

The KMV-Merton model is a revision of the Merton model that varies in certain aspects. It provides an assessment of the likelihood of firms to default based on the ability of the firm equity holders to pay off their debts. Recent studies on the application of the KMV-Merton model include studies on the impact of COVID-19 on default risk [2], the default risk of internet finance companies [3], and the firm's volatility estimation [4]. Other recent studies, such as [5] focused on improving the accuracy of the model. Besides the ability of the model to predict default in advance, there were discussions on the data implementation of the model in capturing enough data on default. This led to many modifications of the KMV-Merton model. It is said that the probability of default from the KMV-Merton model can only be calculated using some comparability analysis based on accounting data [6]. Others said that the KMV-Merton model is not a sufficient statistic for the probability of default [7].

Therefore, in the previous literature, the KMV-Merton model was adapted into a statistical model to form a model to predict the likelihood of firms defaulting. Most studies concluded that the combination of the KMV-Merton model with the statistical models gives better default prediction (see [8]–[13]). From this combination also, the factors that contribute to a certain event were analyzed. For example, one said that the default probability from the KMV-Merton model contributes to the firm's sector analysis [14], and tax arrears predict corporate bank loan defaults [15]. Among the studies, one of the statistical models used is the Discriminant Analysis (DA).

The DA is a statistical technique used to evaluate the differences between each observation. It is the most direct form of a linear combination of utilised variables developed in 1936 by Fisher and can be viewed as predictive [10]. [16] is the pioneer of the univariate approach of DA in bankruptcy prediction. Then, [17] expanded it to a multivariate context and developed the Z-Score model. A variety of research comes from these, making the DA one of the widely used methods in predicting the financial distress of firms [18], [19]. The direct techniques and stepwise approaches are the two most often utilised strategies for developing discriminant models ([20]).

In applying the statistical models, the authors [8]–[10], [12], [13] utilised the same independent variables encompass financial ratios to predict the probability of default of companies. Financial ratios are vital components in analysing the financial distress of firms [21]–[23]. A study found that financial ratios except for cash flow ratios have a significant impact on the firm's probability of default [24]. Meanwhile, [25] found that return on assets, current ratio, debt to total assets ratio, sales to working capital ratio and cash flow to total assets ratio are statistically significant in predicting default. Apart from that, Debt Ratio, Total Assets Turnover, and Net Profit Margin are also found significant to financial distress [26].

Nonetheless, the authors [8]–[10], [12], [13] used different approaches to define the dependent variables. Some used the information from financial statements [8], company rating [9], loan payment [10], the hybrid KMV-Merton and logistic score model [12] and a list of insolvent/distressed companies [13]. This study used a similar approach as [12], where the KMV-Merton default probabilities are used to define the dependent variables. Since we are using the DA, we are not converting the KMV-Merton default probabilities into a logistic score as [12] did.

Accordingly, this paper adapts the KMV-Merton default probabilities to be the dependent variable and selected financial ratios as the independent variables of the DA model to predict the default/non-default of Malaysian publicly listed companies. This study contributes to the obtaining of the discriminant function that can significantly differentiate between default and non-default of the firms. In addition, the probability of default of the selected firms is estimated quarterly using the KMV-Merton model, instead of yearly as mostly done in the previous literature [3], [27], [28]. Lastly, the performance of the discriminant function obtained is evaluated in our study based on Type 1 and Type II errors.

The rest of the paper is organized as follows. In Section 2, the data and the methodology used in this study to predict default and non-default firms are explained. Then in Section 3, the results are presented and discussed accordingly. Lastly, Section 4 concludes.

2. Data and Methodology

This section explains the data, mathematical model and methods used to achieve the goal of this study. This includes an explanation of how default prediction is done using the KMV-Merton model and discriminant analysis.

2.1 Data Description

This study utilised 11-year financial data from six publicly listed Malaysian firms. Half of the firms were rated consistently from AAA to AA- and the other half were rated inconsistently from AA- to D. The data obtained is in the form of quarter data from 2009 to 2019. We used about 70% of the data from 2009 to 2016 as data training to obtain a discriminant function. The discriminant function obtained is then used to predict the default and non-default firms quarterly from 2017 to 2019, using the rest 30% of the data.

The data collected for the default probabilities estimation are market capitalisation obtained from the DataStream, and short-term borrowing and long-term borrowing obtained from the quarterly report of the firms.

The second type of data is for calculating the financial ratios of firms. The data collected for calculating the financial ratios of firms involving current assets, current liabilities, inventory, total liabilities, total assets, shareholders' equity, operating profit, interest expense, account receivables, net sales, and net profit. These quarterly data are obtained from the quarterly reports of the firms.

2.2 Discriminant Analysis

Discriminant analysis (DA) derives a linear combination of independent variables that discriminates between default and non-default firms from an equation that takes the following form [17]:

$$Z = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_N X_N \tag{1}$$

where:
Z is discriminant score,
a_0 is coefficient (discriminant) weight,
$a_1, a_2, \ldots, a_N$ are discriminant coefficients,
$X_1, X_2, \ldots, X_N$ are discriminant variables.

Based on [8], there are two methods widely used for the derivation of the discriminant model, which are the direct and stepwise methods. The difference between the two methods is based on model construction. The direct method is based on model construction, which means that the model is defined in advance and then used in DA. Meanwhile, the stepwise method is a method where a subset of variables is chosen. Five statistical methods can be chosen to undergo stepwise methods: Wilks' lambda [29], Unexplained Variance [30], Mahalanobis distance [31] and Smallest F ratio [32] [33].

Wilks' lambda test can determine whether a link exists between the dependent variable and the explanatory variables [10]. It also tests the importance of the discriminant function by measuring differences between groups [34]. Wilks' lambda test values are always between 0 and 1. A value of 1 indicates that the median is equal, and the most discriminating variable has a lambda value and significance level close to 0. A low lambda value indicates minimal intra-group Variance and, thus, substantial intergroup variation, resulting in a significant difference in class mean.

The value of standardised discriminant function coefficients interprets the discriminant function. It is vital to examine the relative importance of the variables by analysing the absolute magnitude of the coefficient. The higher the standardised coefficient value, the more significant the contribution of the respective variable to the discrimination between groups [35]. This can be interpreted from the results of the standardised canonical discriminant function coefficients. Both studies from [10] and [11] show that the variable that gives the highest discriminant function coefficient is the most crucial variable to its discriminant function.

However, the results of interpretation of the coefficient value did not tell the values of the discriminant function that discriminant between the groups. This is where the group means are important. Group means are called centroids. The different values of the group centroids visualise how the function can discriminate between groups. This is presented by a study from [10] that shows the positive values of the group centroids determine that the firm is healthy, while negative values of the group centroids show that the firm is failing. The value of group centroids depends on the means of each group. Therefore, different studies obtained different values of group centroids.

In summary, this study focuses on obtaining the discriminant function as in Equation (1) for default risk prediction. This study used Wilks' lambda which is a stepwise step to the DA done. Meanwhile, the default probability estimated from the KMV-Merton model is used to determine the Z-Score of Equation (1).

2.3 KMV-Merton's Default Probabilities

The probability of default (*PD*) is defined as the probability that the market value of an asset falls below the face value of debt at time *t*. [36] believed that the dynamic of market value (dV_A) of the underlying properties of the firm follows the Geometric Brownian Motion as follows:

$$d_A = \mu V_A dt + \sigma V_A dW \qquad (2)$$

where V_A is the market value of the firm's asset, μ is the drift rate, σ is the volatility, and *dW* is the Wiener Process. Since the natural log of future asset values is distributed normally, the PD is estimated using the standard normal distribution function of inverse *d* as follows [8]:

$$\text{PD} = NORMS(-d) \qquad (3)$$

for *d* being as distance to default, *d* that is computed by the following equation:

$$d = \left[ln\left(V_{A,t}\right) - ln(B_t) + (\mu - 0.5\sigma^2)T\right]/\sigma\sqrt{T} \qquad (4)$$

where:
B_t is the book value of a firm's liabilities at any time *t* defined as the summation of short-term and half of long-term borrowings. Here, the values of borrowings are the same each day unless there are changes in the quarterly report,

$V_{A,t}$ is the daily market value of an asset at any time *t* defined based on basic accounting definition, firm's market capitalisation plus its book value of liabilities,

μ is the expected firm's asset return calculated by finding the mean of the daily log return of the market value of the asset for each quarter,

σ is the asset volatility that calculated using the standard deviation of the daily log return of the market value of the asset. Then, the standard deviation is multiplied by the square root of the number of trading days, $\sqrt{63}$, to obtain quarterly volatility. The 63 days referred to the number of trading days in each quarter [37],

T is time denoted as one quarter.

2.4 Setting Up the Dependent and Independent Variables of the Discriminant Function

The dependent variable denoted by the Z-score given in Equation (1) is determined based on the PD estimated from the KMV-Merton model. Equations (3) and (4) are used to estimate the PD of the selected firms and are next categorised according to Table 1.

Table 1. Discriminant Score Z

Firm's category	PD	Z-score
Non-default	0-0.49	0
Default	0.50-1.00	1

Meanwhile, the independent variables known by the discriminant variables ($X_1, X_2, X_{3,}\ldots, X_N$) are represented by the financial ratios shown in Table 2.

Table 2. List of financial ratios and formulas [38]

Variables	Financial Ratios	Formula
X_1	Current ratio	Current assets / Current liabilities
X_2	Acid test ratio	(Total current assets-inventory-prepayments) / Current liabilities
X_3	Debt ratio	Total liabilities (total non-current liabilities and current liabilities / Total assets (total non-current assets and current assets)
X_4	Debt to equity ratio	Total liabilities / Shareholders' equity
X_5	Time interest earned ratio	Operating profit / Interest expense
X_6	Total assets turnover	Net sales / Total assets (non-current assets and current assets)
X_7	Account receivables turnover ratio	Trade receivables x 63 / Total net sales
X_8	Net profit margin ratio	Net profit / Total net sales or revenue
X_9	Return on assets ratio	Net profit / Total assets
X_{10}	Return on capital employed ratio	Operating profit / Capital employed

These are three criteria considered in the selection of ratios [9]:
i.   The financial ratios have been identified theoretically as indicators for determining default,
ii. previously used in empirical work to predict insolvency,
iii. and can be computed and determined conveniently from the researcher's database.

2.5 Predicting the Firm's Default Risk

SPSS Discriminant Analysis was employed to run 70% of the data, that is set up as explained in sub-section 2.4 to obtain a discriminant function as in Equation (1). The method used in this study is Wilks' lambda. It is a variable selection method in the stepwise discriminant analysis that chooses variables for the discriminant function. The KMV-Merton model and the discriminant function obtained from running the SPSS discriminant analysis are used to predict the default and non-default firms using the rest of 30% of the data.

The prediction was done before and after the adaptation of the KMV-Merton model into the DA. Therefore, by using Equations (3) and (4), the PD of the firms is estimated and categorised as default or non-default according to Table 1. This represents the prediction before the adaptation was made. The next prediction is after the KMV-Merton model was adapted into the DA model, where the discrimination function Z-Score is obtained as in Equation (1). Here, the discriminant function consists of only significant independent variables. The Z-Score of the firms was calculated and classified into default and non-default groups according to one of the outputs from the SPSS discriminants analysis, which is called functions at the group centroids. Group centroids indicate the mean values for the discriminant functions (Z-score) for a group. Therefore, the Z-Score calculated from the discriminant function that is near the centroid is said to belong to that group.

After that, the performance of each prediction is determined based on Type I and Type II errors. Type I error is defined as incorrectly classified default firm as non-default, while Type II error is defined as incorrectly classified non-default firm as default [39]. All these are expressed in formulae as follows [40]:

$$\text{Accuracy (\%)} = \frac{C_d}{T_d} + \frac{C_s}{T_s} \tag{5}$$

$$\text{Type I error (\%)} = 1 - \frac{C_d}{T_d} \tag{6}$$

$$\text{Type II error (\%)} = 1 - \frac{C_s}{T_s} \tag{7}$$

where T_d is the actual number of default firms, T_s is the actual number of non-default firms, C_d is the number of correctly predicted default firms, and C_s is the number of correctly predicted non-defaulted firms. In actual cases, the default and non-default firms are determined based on the ratings available

for the firms. Specifically, firms with a rating between AAA to BBB- are considered non-default, while B+ to C is considered as default.

3. Results and Discussion

This section discusses the results obtained from the study. It involves a discussion of the results from the data descriptive, discriminant analysis using SPSS, and performance before and after the combination of the KMV-Merton model in the discriminant analysis model.

3.1 Descriptive Statistics

Table 3 shows the data descriptive for estimating firms' default probabilities using the KMV-Merton Model.

Table 3. The data descriptive for the default probabilities estimation

	N	Minimum	Maximum	Mean	Standard Deviation
Market capitalisation	264	73.956	94871.726	15994.516337	23929.833735
Short term borrowing	264	0	14855.141	8584.735807	1870.898180
Long term borrowing	264	0	11386.399	1617.181595	2908.022436

Meanwhile, Table 4 shows the descriptive data of the financial ratios used in the Discriminant Analysis.

Table 4. The data descriptive for the financial ratios

	N	Minimum	Maximum	Mean	Standard Deviation
Current asset	264	10.8461	15109.329	2392.763383	339.8933931
Current liabilities	264	6.3043	10956.308	1724.949362	2757.31698
Inventory	264	.6789	1935.419	148.9530682	265.8803581
Total liabilities	264	44.0111	19652.028	3639.699565	5707.969253
Total assets	264	69.7051	56155.899	9529.74627	16315.067939
Shareholders' equity	264	24.9468	16763.5	2556.981929	3954.369202
Operating profit	264	-1651.89	2040.89	85.23968106	261.7806059
Interest expense	264	-.943	1995.714286	22.83984718	131.70429
Account receivables	264	2.393	5246.37	651.2074602	114.5255567
Net sales	264	2.981	4889.12	547.1006371	993.0591672
Net profit	264	-409.777	1695.973	87.97215492	237.2578048

In Tables 3 and 4, we have 264 samples for each type of financial data of firms. The data with the lowest and highest mean are the interest expense and market capitalization, respectively. Greater standard deviation in the data shows greater variations in the samples. This means the spreads of each data distribution from the mean are all high, with the most found in the data of market capitalization and the least in the data of account receivables.

3.2 Tests of Equality of Group Means and Stepwise Statistics

The purpose of the equality test is to prove the significant differences between non-default and default groups on each of the independent variables. The output from this test can be seen in Table 5. Stepwise statistics show the steps taken for the selection of variables that are included in the analysis. The method chosen for stepwise statistics is Wilks' Lambda. It allows one to select the

variables that will be entered in the discriminant function. The result from this stepwise statistic is given in Table 6.

Table 5. Tests of Equality of Group Means

Discriminant Variables	Wilks' lambda	F	df1	df2	Sig.
X_1	0.994	1.134	1	190	0.288
X_2	0.993	1.398	1	190	0.239
X_3	0.999	0.163	1	190	0.687
X_4	0.993	1.385	1	190	0.241
X_5	1.000	0.015	1	190	0.902
X_6	1.000	0.019	1	190	0.891
X_7	0.992	1.471	1	190	0.227
X_8	0.941	11.982	1	190	0.001
X_9	0.883	25.116	1	190	0.000
X_{10}	0.883	25.112	1	190	0.000

Table 5 shows which ratios are considered the most discriminating variable between the non-default and default groups. It also tests the null hypothesis that the group means are equal across all dependent variables. If Wilk's lambda is smaller than the critical value, then the null hypothesis can be rejected. The variables X_8, X_9, and X_{10} show significant differences in Wilk's lambda and F-values, leading to the rejection of the null hypothesis. These variables present higher values of F's and lower values of Wilks's lambda. The significant value (Sig.) that is close and equal to 0 also indicates that these variables are significant to the discriminant function. This is parallel to Table 6 where both X_8 and X_9 have a high tolerance of 0.931, near 1. It proves that these variables contribute high information to the discriminant function model. Although X_{10} gives a high correlation to the discriminant function, we found that X_{10} and X_9 are detected with multicollinearity in the pooled within-groups matrices test. Therefore, one of the variables must be removed to avoid potential problems in the prediction.

Table 6. Variables in the Analysis

Step		Tolerance	F to Remove	Wilks' Lambda
1	X_9	1.000	25.116	
2	X_9	.931	16.918	.883
	X_8	.931	4.345	.941

3.3 Test of Homogeneity of Covariance Matrices

Tests of homogeneity of covariance matrices include the outputs from Tables 7 and 8. This test is conducted to show whether the covariance matrices are equivalent or not. It can be evaluated through the null hypothesis.

Table 7. Log Determinants

Default Probability	Rank	Log Determinant
Non-default	2	20.062
Default	2	22.781
Pooled within-groups	2	20.233

Table 8. Test Results

Box's M		21.683
F	Approx.	5.883
	df1	3
	df2	553.242
	Sig.	0.001

Table 7 presents the log determinants for the group's covariance matrix and the pooled within-group covariance. The Default group shows the highest log determinant value, indicating a difference in its covariance matrix. This is supported by Table 8. Table 8 tests the null hypothesis that the covariance matrices do not differ between groups formed by the dependent variables. The null hypothesis is that if the significance is greater than 0.05, the covariance matrices are equal (H0), and if the significance is less than 0.05, the covariance matrices are not equal (H1). Table 8 shows the high value of Box's M, which is 21.683, and the significance (Sig.) of F tends to be 0, which is less than 0.05. According to [10], the Box's M values must be high, and the significance of the F must be near 0 for the analysis to be valid, which is to be unequal in covariance matrices. Thus, this study rejects the null hypothesis of equal covariance matrices. In addition, the amount of "Rank" in Table 7 represents two significant independent variables that can be used for the discriminant function model.

3.4 Summary of Canonical Discriminant Function

In summary of the canonical discriminant function, five outputs are obtained, and they are presented in Tables 9-14.

Table 9. Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	0.158	100.0	100.0	0.370

Note: Function 1 canonical discriminant functions were used in the analysis.

Table 9 provides information on the discriminant functions produced. Eigenvalue indicates the measure of association between the discriminant function and the dependent variable. A higher eigenvalue (near 1) displays a stronger discriminant function model. Like the canonical correlation, the value that is near to one presents a better discriminant function model [10]. In this study, the canonical correlation is 0.370, and the eigenvalue of 0.158, which is not extremely high. However, this model is statistically significant based on Table 10. Hence, it can still be considered a good model [11].

Table 10. Wilks' lambda

Test of Function(s)	Wilks' lambda	Chi-square	df	Sig.
1	0.863	27.762	2	0.000

Table 10 shows Wilks' lambda of the discriminant function. The closer Wilks' lambda value to 0 illustrates the higher quality of the model [10]. Here, the discriminant function model could be better based on the Wilks' lambda 0.863. However, the discriminant function is said to be able to discriminate the groups based on the significant value (Sig.) at 0.000.

Table 11. Structure Matrix

Discriminant Variables	Function 1
X_9	.914
$X_{10}{}^a$	.913
X_8	.631
$X_4{}^a$	.387
$X_7{}^a$	-.212
$X_3{}^a$	-.159
$X_1{}^a$	-.046
$X_6{}^a$	.019
$X_5{}^a$	-.015
$X_2{}^a$	-.012

Note: Pooled within-group correlations between discriminating variables and standardised canonical discriminant functions. Variables are ordered by the absolute size of correlation within the function. (a) indicates the variables that are not used in the analysis.

Table 11 provides the structure matrix, which illustrates the importance of correlations between discriminant variables and discriminant function. By ignoring the variables that are not used in the analysis, X_9 presents the highest absolute size of correlation within functions, followed by X_8. Hence, X_9 is considered the most important variable in determining the firm's default risk, followed by X_8. Thus, it is confirmed that X_8 (net profit margin ratio) and X_9 (return on assets ratio) are suitable independent variables for the discriminant function of the model.

Table 12. Canonical Discriminant Function Coefficients

Selected discriminant variables	Function 1
X_8	.009
X_9	.001
(Constant)	-.301

Note: Unstandardised coefficients.

Table 12 presents the list of coefficients of the independent variables X_8 and X_9. These are used to form an unstandardised discriminant function expressed as:

$$Z = -0.301 + X_8 * 0.009 + X_9 * 0.001 \qquad (8)$$

Equation (8) represents the discriminant function Z-score, which is the result of the adaptation of the KMV-Merton model and financial ratios into the DA. It is used in this study to make predictions on default and non-default groups.

Table 13. Functions at Group Centroids

Default Probability	Function 1
Non-default	-.065
Default	2.420

Note: Unstandardised canonical discriminant functions evaluated at group means.

Table 13 displays the average Z value of the discriminant function. It indicates that the value of functions at group centroids for non-default firms is at -0.065, and the centroids for default firms is at 2.420. Therefore, when the Z-value in Equation (8) of the firm is negative, the firm is predicted to be a non-default firm, while when the Z-value obtained is positive, the firm is predicted to default.

Table 14. Classification Results

			Predicted Group Membership		Total
			Non-default	Default	
Original[a]	Count	Non-default	174	13	187
		Default	2	3	5
	%	Non-default	93.0	7.0	100.0
		Default	40.0	60.0	100.0
Cross-validated[b,c]	Count	Non-default	69	1	70
		Default	2	0	2
	%	Non-default	98.57	1.43	100
		Default	100	0	100

a. 92.2% of original grouped cases were correctly classified.
b. Cross-validation is done only for those cases in the analysis. In cross-validation, each case is classified by the functions derived from all cases other than that case.
c. 92.2% of cross-validated grouped cases are correctly classified.

Table 14 summaries the performance of the discriminant function (8) to classify the two groups, default, and non-default firms, based on the trained data. Out of 187 cases, 174 were correctly predicted as non-default, and 13 were incorrectly predicted. In the meantime, only three were correctly predicted for the default group out of 5. This produces a higher accuracy of about 93% for the non-default group compared to 60.0% for the default group.

3.5 Default Risk Prediction

The default and non-default of the selected firms are predicted at first using the KMV-Merton model and next using the discriminant function model in (8). This discriminant function represents the adaptation of the KMV-Merton model into the DA. Afterwards, the performance of each prediction is compared according to the percentage of accuracy, Type I and Type II errors calculated based on Equations (5), (6), and (7). The results are presented in Tables 15 and 16.

Table 15. Prediction Results

Model			Predicted		Total
			Non-default	Default	
Discriminant Function	Actual	Non-default	**47**	1	48
		Default	22	**2**	24
KMV-Merton	Actual	Non-default	**46**	2	48
		Default	24	**0**	24

Table 16. The accuracy, Type I, and Type II errors

	KMV-Merton Model	Discriminant Function
Accuracy (%)	63.89	68.06
Type I error (%)	100	91.67
Type II error (%)	4.17	2.08

Tables 15 and 16 show that both the KMV-Merton model and the discriminant function can predict the non-default group well compared to the default group. This is shown as the percentage of incorrectly predicted for the non-default group (Type II error) is extremely low than the default group (Type I error). This is contrary to the study done by [39] where Type I error is found lower than Type II error. Higher Type I error is due to the unequal data between the non-default and default groups that lead to a potential bias in the data-trained model, as shown in Table 14. Therefore, the discriminant function tends to be more biased in predicting the non-default group and causes high errors in the default group prediction. Overall, the accuracy for both the KMV-Merton model and the discriminant function is low. However, the discriminant function gives higher accuracy in prediction

than the KMV-Merton model. This means that the addition of two financial ratios (X_8 and X_9) into a discriminant function may improve the KMV-Merton model performance to predict default risk.

4. Conclusion

Several studies have been conducted previously in the field of firm default prediction. In this study, the non-default and default firms are predicted quarterly. It gives a more detailed focus on the firm's financial performance as it evaluates through quarters. Essentially, the default prediction is done based on the combination of financial ratios and default probability of the KMV-Merton model in the discriminant analysis (DA) model.

This combination leads to a discriminant function with net profit margin (NPM) and return on asset (ROA) as its determinant variables. NPM and ROA are found significant (p-value = 0) in discriminating the default/non-default of the firms with Wilks's lambda = 0.941, F(1,190) = 11.982 for NPM and Wilks's lambda = 0.883, F(1,190) = 25.116 for ROA. Although the return on capital employed ratio is also found significant, it is detected with multicollinearity in the pooled within-groups matrices test.

In the discriminant analysis, the discriminant function failed to classify the default group well due to the 60% accuracy compared to the 93% accuracy of classifying the non-default group. This may be due to the limited and imbalanced data on default firms in the data training process. This makes the discriminant analysis model biased in making predictions to a majority non-default group. Plus, the eigenvalue and canonical correlation of the discriminant function are low at 0.158 and 0.370, respectively. However, the discriminant function could be better as it can discriminate the groups based on Wilks' lambda = 0.863 and p-value at 0.000.

In addition, we found that the accuracy of predicting default using the combination of the KMV-Merton model and financial ratios in the DA (68% accuracy) is slightly higher than using the KMV-Merton model alone (63% accuracy). This is also supported by the lower values of Type 1 and Type II errors for the discriminant functions compared to the KMV-Merton model. Hence, the adaptation of KMV-Merton's default probabilities and financial ratios in the DA model is said to be able to improve the firms' default prediction.

In future research, the model could be improved as more data on default firms are included in the samples. Nevertheless, data acquisition on default firms has been limited due to several factors. In the alternative, the group classification may be redefined in future works to obtain an equal distribution for each group.

Conflict of Interest
The authors declare no conflict of interest in the subject matter or materials discussed in this manuscript.

References

[1] M. S. Tosun and S. Yildiz, "How Does Aggregate Tax Policy Uncertainty Affect Default Risk?," *J. Risk Financ. Manag.*, vol. 13, no. 12, pp. 1–17, 2020, doi: 10.3390/jrfm13120319.

[2] J. Welburn and A. Strong, "Estimating the Impact of COVID-19 on Corporate Default Risk," 2020. doi: 10.7249/wra173-2.

[3] L. Zeng, W. Lau, and E. N. A. Bahri, "Can the Modified ESG-KMV Logit Model Explain the Default Risk of Internet Finance Companies?," *Front. Environ. Sci.*, vol. 10, 2022, doi: 10.3389/fenvs.2022.961239.

[4] B. Christoffersen, D. Lando, and S. B. Nielsen, "Estimating Volatility in the Merton Model: The KMV Estimate Is Not Maximum Likelihood," *Math. Financ.*, vol. 32, pp. 1214–1230, 2022, doi: 10.1111/mafi.12362.

[5] X. Cheng, Z. Sun, and W. Bao, "Study on credit risk of real estate industry based on genetic algorithm KMV model," *J. Phys. Conf. Ser.*, vol. 1629, no. 1, p. 12072, 2020, doi: 10.1088/1742-6596/1629/1/012072.

[6] S. Anthony and L. Allen, *Credit risk measurement: new approaches to value at risk and other paradigms(google pdf book)*. New York: John Wiley & Sons, Inc., 2002.

[7]   S. T. Bharath and T. Shumway, "Forecasting Default with the Merton Distance to Default Model," *Rev. Financ. Stud.*, vol. 21, no. 3, pp. 1339–1369, May 2008, doi: 10.1093/rfs/hhn044.

[8]   S. Aktan, "Early Warning System for Bankruptcy: Bankruptcy Prediction," *Thesis*, pp. 1–188, 2011.

[9]   V. Gupta, "An Empirical Analysis of Default Risk for Listed Companies in India: A Comparison of Two Prediction Models," *Int. J. Bus. Manag.*, vol. 9, no. 9, pp. 223–234, 2014, doi: 10.5539/ijbm.v9n9p223.

[10]   Z. Khemais, D. Nesrine, and M. Mohamed, "Credit Scoring and Default Risk Prediction: A Comparative Study between Discriminant Analysis & Logistic Regression," *Int. J. Econ. Financ.*, vol. 8, no. 4, p. 39, 2016, doi: 10.5539/ijef.v8n4p39.

[11]   N. T. Duong, D. Thi Thu Ha, and N. Bich Ngoc, "The Application of Discriminant Model in Managing Credit Risk for Consumer Loans in Vietnamese Commercial Bank," *Asian Soc. Sci.*, vol. 13, no. 2, p. 176, 2017, doi: 10.5539/ass.v13n2p176.

[12]   P. Andrikopoulos and A. Khorasgani, "Predicting unlisted SMEs' default: Incorporating market information on accounting-based models for improved accuracy," *Br. Account. Rev.*, vol. 50, no. 5, pp. 559–573, 2018, doi: https://doi.org/10.1016/j.bar.2018.02.003.

[13]   P. Gaba, S. Arora, A. Lall, V. Khurana, and A. Rathi, "Forecasting and Comparing the Results From Altman Z - Score and Merton Distance To Default Model in the Indian Scenario," *Int. J. Adv. Res.*, vol. 7, no. 10, pp. 1295–1309, 2019, doi: 10.21474/ijar01/9959.

[14]   V. A. Delapedra-Silva, "The Bankruptcy Risk in Infrastructure Sectors: An Analysis From 2006 to 2018," *Ram Rev. Adm. Mackenzie*, vol. 22, no. 4, 2021, doi: 10.1590/1678-6971/eramf210104.

[15]   K. Kohv and O. Lukason, "What Best Predicts Corporate Bank Loan Defaults? An Analysis of Three Different Variable Domains," *Risks*, vol. 9, no. 29, 2021, doi: 10.3390/risks9020029.

[16]   W. H. Beaver, "Financial Ratios As Predictors of Failure," *J. Account. Res.*, vol. 4, pp. 71–111, Apr. 1966, doi: 10.2307/2490171.

[17]   E. I. Altman, "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy," *J. Finance*, vol. 23, no. 4, pp. 589–609, Apr. 1968, doi: 10.2307/2978933.

[18]   I. Wahyudi, I. Hakiman, and T. Mm, "Analysis of Factors to Predict Financial Distress," *Int. J. Res. Stud. Publ.*, vol. 10, no. 7, 2020, doi: 10.29322/ijsrp.10.07.2020.p10372.

[19]   J. Lord, A. Y. Landry, G. T. Savage, and R. Weech-Maldonado, "Predicting Nursing Home Financial Distress Using the Altman Z-Score," *Inq. J. Heal. Care Organ. Provis. Financ.*, vol. 57, 2020, doi: 10.1177/0046958020934946.

[20]   F. Sun, Y. Chen, Y. Qiu, S. Wang, and S. Liang, "Systematic vs. Stepwise Parameter Optimization for Discriminant Model Development: A Case Study of Differentiating Pinellia Ternata From Pinellia Pedatisecta With Near Infrared Spectroscopy," *J. Near Infrared Spectrosc.*, vol. 28, no. 5–6, pp. 287–297, 2020, doi 10.1177/0967033520924579.

[21]   A. K. Benson, T. Swanepoel, and H. J. van Vuuren, "Firm's Value Sustainability via Accounting Ratios: The Case of Nigerian Listed Firms," *J. Econ. Financ. Sci.*, vol. 14, no. 1, 2021, doi: 10.4102/jef.v14i1.529.

[22]   İ. E. Ceylan, "The Impact of Firm-Specific and Macroeconomic Factors on Financial Distress Risk: A Case Study From Turkey," *Univers. J. Account. Financ.*, vol. 9, no. 3, pp. 506–517, 2021, doi: 10.13189/ujaf.2021.090325.

[23]   Y. Tong and Z. Serrasqueiro, "Predictions of Failure and Financial Distress: A Study on Portuguese High and Medium-High Technology Small and Mid-Sized Enterprises," *J. Int. Stud.*, vol. 14, no. 2, 2021, doi: 10.14254/2071-8330.2021/14-2/1.

[24]   A. M. A. Bajwa and M. H. Rashid, "Financial Ratios : A Tool for Computing Probability of Corporate Default," *Audit Account. Rev.*, vol. 1, no. 2, 2021, doi 10.32350/aar.12.04.

[25]   H. S. Shetty and T. N. Vincent, "Corporate Default Prediction Model: Evidence From the Indian Industrial Sector," *Vis. J. Bus. Perspect.*, vol. 28, no. 3, 2021, doi: 10.1177/09722629211036207.

[26]   H. Herbowo and E. R. Saputri, "Pengujian Rasio Keuangan Terhadap Financial Distress: Bukti Di Indonesia," *Wahana J. Ekon. Manaj. Dan Akunt.*, vol. 26, no. 1, pp. 1–15, 2023, doi: 10.35591/wahana.v26i1.805.

[27]   D. Malasari, M. Adam, Y. Yuliani, and A. Hanafi, "Financial Ratios and Probability of Default by Using The KMV-Merton Method in the Non-Financial Sector Listed on the Indonesia Stock Exchange," *Financ. Theory Pract.*, vol. 24, no. 1, 2020, doi: 10.26794/2587-5671-2020-24-1-6-13.

[28]   M. Mushafiq, S. A. Sami, M. K. Sohail, and M. I. Sindhu, "Merton-Type Default Risk and

Financial Performance: The Dynamic Panel Moderation of Firm Size," *J. Econ. Adm. Sci.*, vol. 40, no. 2, pp. 168–181, 2022, doi: 10.1108/jeas-09-2021-0181.

[29] X. Bai and Z. Zhao, "An Optimal Credit Scoring Model Based on the Maximum Default Identification Ability for Chinese Small Business," *Discret. Dyn. Nat. Soc.*, vol. 2022, pp. 1–14, 2022, doi: 10.1155/2022/1551937.

[30] R. B. Avery, P. S. Calem, and G. B. Canner, "Consumer credit scoring: Do situational circumstances matter?," *J. Bank. Financ.*, vol. 28, no. 4, pp. 835–856, 2004, doi: 10.1016/S0378-4266(03)00202-4.

[31] M. Rezapour, "Anomaly Detection using Unsupervised Methods : Credit Card Fraud Case Study," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 11, pp. 1–8, 2019.

[32] V. B. Djeundje, J. Crook, R. Calabrese, and M. Hamid, "Enhancing credit scoring with alternative data," *Expert Syst. Appl.*, vol. 163, p. 113766, 2021, doi: https://doi.org/10.1016/j.eswa.2020.113766.

[33] N. M. Yusof, I. Q. Alias, A. J. M. Kassim, and F. L. N. M. Zaidi, "Determining the credit score and credit rating of firms using the combination of KMV-Merton model and financial ratios," *Sci. Technol. Indones.*, vol. 6, no. 3, pp. 105–112, 2021, doi: 10.26554/sti.2021.6.3.105-112.

[34] M. Şahin and Ş. Koç, "A Monte Carlo Simulation Study Robustness of MANOVA Test Statistics in Bernoulli Distribution," *Süleyman Demirel Univ. J. Nat. Appl. Sci.*, vol. 22, no. 3, pp. 1125–1131, 2018, [Online].

[35] E. Puskarczyk, "Application of multivariate statistical methods and artificial neural network for facies analysis from well logs data: An example of Miocene deposits," *Energies*, vol. 13, no. 7, pp. 1–18, 2020, doi: 10.3390/en13071548.

[36] R. C. Merton, "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates," *J. Finance*, vol. 29, no. 2, pp. 449–470, Apr. 1974, doi: 10.2307/2978814.

[37] W. Chun-Ping, Q. Lin-Jing, and L. Jian-Wei, "Commercial bank credit risk measurement based on KMV model studies," in *Proceedings of the 2015 International Conference on Engineering Management, Engineering Education and Information Technology*, 2015, pp. 456–465.

[38] D. A. W. Akma Hidayu, A. J. Ja'izah, and W. T. Wan Mardyatul Miza, *Financial and Management Accounting*, 2nd ed. Malaysia: McGraw-Hill Education Malaysia, 2016.

[39] Y. Zizi, A. J. Alaoui, B. El Goumi, M. Oudgou, and A. El Moudden, "An Optimal Model of Financial Distress Prediction: A Comparative Study Between Neural Networks and Logistic Regression," *Risks*, vol. 9, no. 11, p. 200, 2021, doi: 10.3390/risks9110200.

[40] M. S. Çolak, "A New Index Score for the Assessment of Firm Financial Risks," Turkey, 2019.

Biography of all authors

Picture	Biography	Authorship contribution
	Nur Ain Al-Hameefatul Jamaliyatul is currently a master's student in the field of mathematical sciences at the College of Computing, Informatics, and Mathematics, Universiti Teknologi MARA, Johor Branch, Segamat Campus.	Conducting a research and investigation process, specifically collecting, implementing, analysing, and interpreting the data. Also, writing the initial draft of the article.
	Nurul Afiqah Zainuddin is currently working in the industrial field after completing her degree in mathematics at the College of Computing, Informatics, and Mathematics, Universiti Teknologi MARA (UiTM) Negeri Sembilan Branch, Seremban Campus.	Conducting a research and investigation process, specifically collecting, implementing, analysing, and interpreting the data. Also, writing the initial draft of the article.
	Izza Suraya Zulhazmi is currently working in the industrial field after completing her degree in mathematics at the College of Computing, Informatics, and Mathematics, Universiti Teknologi MARA (UiTM) Negeri Sembilan Branch, Seremban Campus.	Conducting a research and investigation process, specifically collecting, implementing, analysing, and interpreting the data. Also, writing the initial draft of the article.
	Norliza Muhamad Yusof is a senior lecturer at the College of Computing, Informatics, and Mathematics, Universiti Teknologi MARA (UiTM) Negeri Sembilan Branch, Seremban Campus. Most of her studies are in the field of financial mathematics, focusing on default risk measurement.	Design the research work and write the article.
	Muhamad Luqman Sapini is a senior lecturer at the College of Computing, Informatics, and Mathematics, Universiti Teknologi MARA (UiTM) Negeri Sembilan Branch, Seremban Campus. Most of his studies are in the field of applied mathematics, focusing on Dynamical Systems and Topological Data Analysis.	Review and edit the article.

Comparative Analysis of Euler and Runge-Kutta Fehlberg Methods in Solving the Lotka-Volterra Competitive Model

Nur Hidayah Abd Rahaman
College of Computing, Informatics, and Mathematics, Universiti Teknologi MARA, Perlis Branch Arau
Campus, Perlis, Malaysia
2020878092@student.uitm.edu.my

Nurizatul Syarfinas Ahmad Bakhtiar
College of Computing, Informatics, and Mathematics, Universiti Teknologi MARA, Perlis Branch Arau
Campus, Perlis, Malaysia
nurizatul@uitm.edu.my

Hafizah Hajimia
College of Arts and Sciences Education and Modern Languages, Universiti Utara Malaysia,
Kedah, Malaysia
hafizah.hajimia@uum.edu.my

Nur Fatihah Fauzi
College of Computing, Informatics, and Mathematics, Universiti Teknologi MARA, Perlis Branch Arau
Campus, Perlis, Malaysia
fatihah@uitm.edu.my

Nur Izzati Khairudin
College of Computing, Informatics, and Mathematics, Universiti Teknologi MARA, Perlis Branch Arau
Campus, Perlis, Malaysia
zat.khairudin@uitm.edu.my

Article Info	ABSTRACT
	This study addresses the problem of determining equilibrium and stability in ecosystems where species compete for the same resources, as outlined by the principle of competitive exclusion. It is often challenging to ascertain the rate at which exclusion happens and whether species can coexist in competitive environments. To discover these issues, the study investigates the competitive interactions between lions (Panthera Leo) and leopards (Panthera pardus) in the Sabi Sand Game Reserve, South Africa, focusing on whether the dominant competitor, lions, limit the population and distribution of leopards. Using the Lotka-Volterra Competitive model, the research compares numerical solutions obtained through Euler and Runge-Kutta Fehlberg (RKF) methods. It also examines how carrying capacity and initial conditions influence equilibrium and stability in this competition. Data on dietary overlaps between lions and leopards were used to test their competitive dynamics, with lions targeting larger prey and leopards focusing on smaller prey. The findings indicate that the RKF method provides more accurate approximations than Euler's method, with leopards showing a higher carrying capacity and greater resilience in the face of competition from lions. This indicates that leopard populations are less affected by lion presence. The study emphasizes the role of carrying capacity in species survival during competition and highlights the utility of numerical methods for predicting competition outcomes without extended experimentation. These findings contribute to wildlife management strategies, particularly in efforts to restore large carnivores in ecosystems, and improve understanding of competitive exclusion in ecological systems.

Corresponding Author:
Name: Nurizatul Syarfinas Ahmad Bakhtiar
Affiliation: College of Computing, Informatics, and Mathematics, Universiti Teknologi MARA, Perlis Branch
Arau Campus, Perlis, Malaysia
Email: nurizatul@uitm.edu.my

1. Introduction

The Lotka-Volterra equation is used to interpret population dynamic in which two organisms interact in one of two ways, either competing for shared resources or associated with a prey-predator system. This study focuses on competitive interaction between two different species solved by Competitive Lotka-Volterra model. Species interactions can be classified into two distinct groups whether interspecific interactions or intraspecific interactions. Interspecific interactions are interactions between two species while intraspecific refers to the interactions between two individuals of the same species. The Competitive Lotka-Volterra equation is a simple model used to solve population dynamics of species competing for shared resources. Competition can be defined as a direct or indirect interaction of organisms that leads to changes in fitness when organisms share the same resources [1].

If natural resources are insufficient to sustain both populations, population growth and survival may be impacted [2]. Consequently, an ecosystem's population may be wiped out. There are equations that are often very difficult to solve analytically. Numerical methods can be used to derive approximate solutions to process models for which analytical solutions are unavailable [3]. Therefore, this study systematically analyzes and compares the exact solution of the numerical solutions obtained by the Euler and Runge-Kutta Fehlberg (RKF) methods on the Lotka-Volterra Competitive model. These two techniques will be compared to determine which method solves this model most effectively.

2. Literature Review

The Lotka-Volterra competition model is a set of mathematical equations used to describe the dynamics of two species that are competing for the same limited resources in an ecological community. These equations describe how the populations of the two species change over time in response to their interactions. According to Seytov *et al.* [4], Lemos-Silva *et al.* [5], Akjouj *et al.* [6], the Lotka-Volterra dynamical system is a set of two autonomous and nonlinear differential equations that describe the interaction between two species - the prey and the predator - and how it influences the growth of both populations. The system is important for studying population dynamics because it provides a mathematical model for understanding the dynamics of prey-predator relationships in an ecosystem.

In this study, two numerical methods, the Euler Method and the Runge-Kutta-Fehlberg (RKF) Method, are applied to solve the model. A numerical solution approximates the solution of a mathematical equation. It is often used where analytical solutions are hard or impossible to get. Euler's method is a first-order numerical procedure for solving ordinary differential equations (ODEs) with a given initial value. It is a basic explicit method for numerical integration of ODEs [7]. It is the first numerical method for solving initial value problem (IVP) and serves to illustrate the concepts involved in the advanced methods. The RKF method is an algorithm in numerical analysis for solving ODEs with adaptive step sizes to balance accuracy. The RKF method is derived from the calculation of two Runge-Kutta (RK) methods of a different order, where subtracting the results from each other can obtain an estimate of the error [8], [9].

Recently, many researchers applied numerical experiments using different schemes to explore competition of dynamics and coexistence among biological populations [10-12]. Razali *et al.* [13] employed various numerical methods, such as Euler, Taylor Series, and Runge-Kutta (RK) methods, to understand the effect of interspecific competition, and found that the RK method provided the most accurate approximation of the orbit's behaviour. Paul *et al.* [14] conducted a comparison between the RKF method and the Laplace Adomian Decomposition Method (LADM) applied to the Lotka-Volterra model. The results indicated that the RKF method exhibited higher accuracy and reliability in solving differential equation models related to population dynamics.

Previous research by Chandra *et al.* [15], Han [16] has used the methodology known as Heun's and Euler's method, which incrementally approximates the solution to two differential equations using first-order derivatives starting from initial conditions. The increments used are small

96

to ensure high accuracy. Euler's method is a first-order numerical procedure for approximating differential equations given an initial value.

Paul *et al.* [8] stated that RKF method could be used as a numerical approximation in a wide range of deterministic and stochastic, linear and nonlinear problems in physics, biology, and chemistry [17] conducted a numerical experiment to illustrate the competitive dynamics between two species using an invasive difference scheme and found that strong competition between two populations with low growth rates and large diffusion coefficients would lead to the extinction of the weaker population. studied the asymptotic behavior, and the number of coexistence equilibria is shown by a saddle-node bifurcation of the level of resource under conditions on competitive effects relative to the associated growth rate per unit of resource. Analysis in this research proved that the key factor of the competition outcome is the relation between intraspecies and interspecies interference effects and sometimes the resource level importance. Extinction in the Lotka-Volterra model has been studied to show the extinction time scale has a power law which depends on the population size. Previous research considers two herbivorous species, rabbits and sheep, competing over a limited amount of food supply [18].

This research's main objective is to assess and compare the effectiveness of two numerical methods, the Euler and RKF methods, in solving the Lotka-Volterra competitive model. The research focuses on the application of these numerical techniques to analyze mathematical models like the Lotka-Volterra competitive model. To conduct this analysis, the researchers will draw upon data from a study that explored competitive interactions between two prominent top predators, namely lions *(Panthera leo)* and leopards *(Panthera pardus)* observed in the Sabi Sand Game Reserve, South Africa, spanning the years 2010 to 2015 [19]. The emphasis will be on determining the species' stability and equilibrium, and it will also determine whether the species can achieve a stable equilibrium or undergo competitive exclusion.

3. Methodology

This study used data from Manaf *et al.* [9] to examine competitive interactions between two top predators, lions *(Panthera Leo)* and leopards *(Panthera Pardus).* Nutritional data obtained concurrently on the two species were used to determine if lions, as the dominant competitor, would limit the distribution and abundance of leopards. The initial populations of both predator species were 254 lions and 355 leopards, and the population growth was observed based on monthly counts of individual lions and leopards known to be alive in the study area. The Euler and Runge-Kutta-Fehlberg (RKF) methods will be applied for solving the Lotka-Volterra Competitive model and subsequently results will be compared with the exact solutions.

3.1 Exact Solution

Growth of population with logistic equation from Lotka-Volterra model are given by:

$$\frac{dx}{dt} = rx\left(1 - \frac{x}{K}\right)$$

(1)

where $x(t)$ is the mean density (in individuals) at time t (in months), r is the instantaneous rate of increase (birth/deaths), and K is the carrying capacity. Assume constant K and r, linear density dependence, no time lags, no migration, no age structure, and limited resources. This equation is to determine K and r values with simple iterative process for lions and leopards.

The solution for the initial condition can be determined by solving equation (1), which gives:

$$x = \frac{K}{1 + \frac{1}{69}e^{(-rt)}(K - 69)}$$

(2)

to utilize equation (2), values for K and r can be determined through a process of curve fitting by referring Figure 1 and Figure 2. The fit is reasonable for:

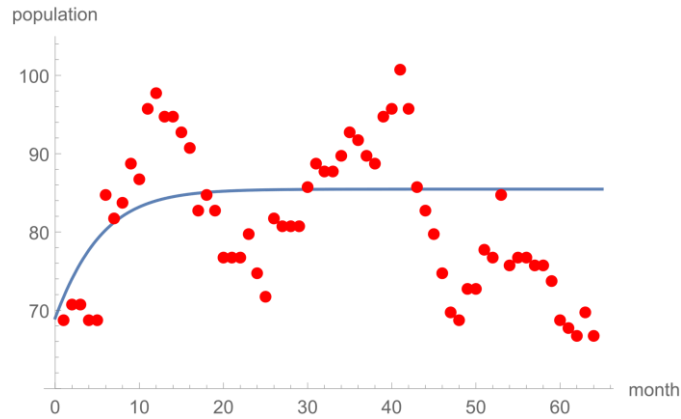Species 1 is lions *(Panthera Leo)* where $K = 85.4648$ and $r = 0.2168$:

Figure 1. Curve Fitting for growth of lions

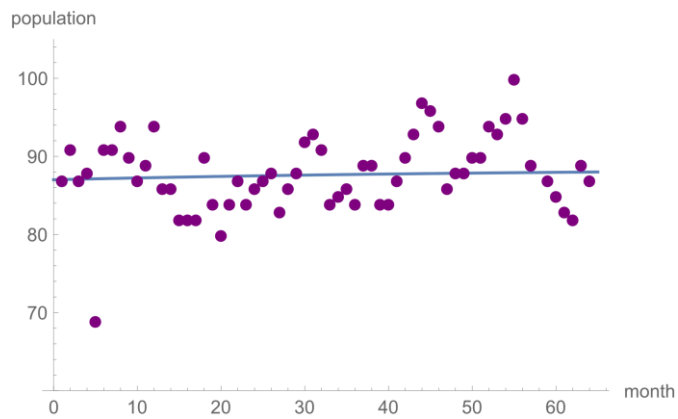Species 2 is leopards *(Panthera Pardus)* where $K = 88.4648$ and $r = 0.0177$:

Figure 2. Curve fitting for growth of leopards

By analyzing Figures 1 and 2, values of r and K for both species can be obtained through the application of the best fit of the fitting curve. This process is facilitated using Mathematica 13.2 software. The values of r and K will then be utilized to compute the exact solution using Equation (2). Subsequently, these determined values will be inserted into Equation (1) to implement the numerical methods, namely the Euler and RKF methods.

3.2 Euler Method
Euler method is the simplest one-step method which uses the concept of local linearity to join multiple small line segments so that they make up an approximation of the actual curve of y versus t. Euler formula can be written as:

$$y_1 = y_0 + hf(x_0, y_0) \tag{3}$$

with the initial values are $x_0 = 0$ and $y_0 = 0$. The step size is $h = 1$ as it represents observation time. Data are taken in an average per month, every month through almost 6 years from 0 until 64 months.

3.3 Runge-Kutta-Fehlberg (RKF) Method
RKF method is a one-step algorithm method with an adaptive step size which automatically organizes the step size as a recompositing to the calculation truncation errors [5].

Consider the following system of ith equation with the initial value problem:

$$y'(t) = f_i(t, y_1, ..., y_n);$$
$$y_i(t_0) = y_0$$

(4)

where i = 1, 2...n.

The equation solves the initial value problem using RK methods of order 4 and order 5. By the first definition:

$$k_1 = hf_i(t, y_i),$$

$$k_2 = hf_i(t + \frac{1}{4}h, y_i + \frac{1}{4}k_1),$$

$$k_3 = hf_i(t_i + \frac{3}{8}h, y_i + \frac{3}{32}k_1 + \frac{9}{32}k_2),$$

$$k_4 = hf_i(t_i + \frac{12}{13}h, y_i + \frac{1932}{2197}k_1 + \frac{7200}{2197}k_2 + \frac{7296}{2197}k_3),$$

(5)

$$k_5 = hf_i(t_i + h, y_i + \frac{439}{216}k_1 + 8k_2 + \frac{3680}{513}k_3 - \frac{845}{4104}k_4),$$

$$k_6 = hf_i(t_i + h, y_i - \frac{8}{27}k_1 + 2k_2 - \frac{3544}{2565}k_3 + \frac{1859}{4104}k_4 - \frac{11}{40}k_5),$$

where h is step size, t_i is time, and $k_1, k_2, ..., k_6$ is every step of accuracy test.
Then, an approximation to the solution of initial value problem is made using RK method of order 4:

$$y_{i+1} = y_i + \frac{25}{216}k_1 + \frac{1408}{2565}k_3 + \frac{2197}{4101}k_4 - \frac{1}{5}k_5$$

(6)

A better value for the solution is determined using RK method of order 5:

$$z_{i+1} = y_i + \frac{16}{135}k_1 + \frac{6656}{12825}k_3 + \frac{28561}{56430}k_4 - \frac{9}{50}k_5 + \frac{2}{55}k_6$$

(7)

Calculation is repeated by using those values.

3.4 Lotka-Volterra Competitive Model

The Lotka-Volterra competitive model is a set of equations that defines how two species interact in a competitive environment. Research from Manaf [9] stated that the Lotka-Volterra model of interspecific competition is based on two other models of population growth, namely the exponential growth and logistic sigmoid growth models.

$$\frac{dN_1}{dt} = r_1 N_1(t) \left[\frac{k_1 - N_1(t) - \beta_{12} N_2(t)}{k_1} \right]$$

$$\frac{dN_2}{dt} = r_2 N_2(t) \left[\frac{k_2 - N_2(t) - \beta_{21} N_1(t)}{k_2} \right]$$

(8)

where, N represents the population density of species for lions and leopards. The term r represents the instantaneous rate of increase of both species and K represents the carrying capacity of the species. The parameter β_{21} represents the per capita effect of lions on the population growth of leopards' species and β_{12} represents the per capita effect of leopards on the population growth of lions. Equation (8) must be solved numerically using the Mathematica 13.2 software because an exact analytical solution would take a long time to obtain.

.

4. Results and Discussion

Logistic equations for the population of lions and leopards tested using numerical approximation methods (Euler and RKF methods) are shown as below. Based on Figures 3 and 4, the graph shows that RKF method values are more accurate than Euler method values because the RKF curve is closer to the exact solution curve. For solving the Lotka-Volterra competitive model, it was found that the RKF method is more reliable than the Euler method and this is applicable to both species.
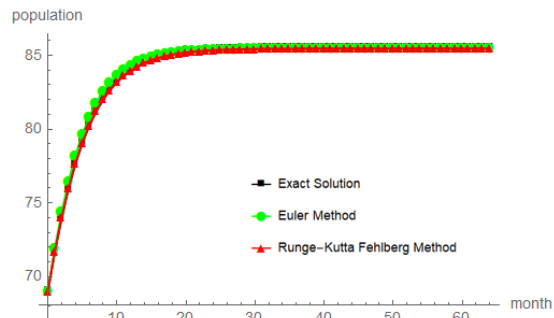

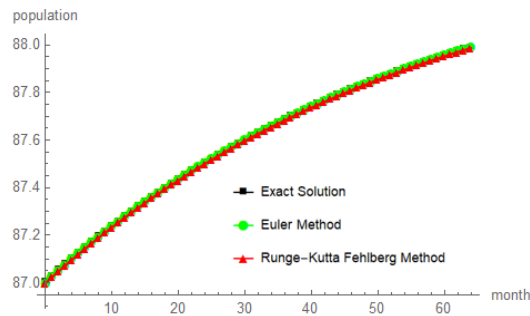Figure 3. Numerical approximation graphs for lions' logistic curve

Figure 4. Numerical approximation graphs for leopards' logistic curve

Table 1. The Mean Absolute Error (MAE) for lion and leopard

Method	Euler	Runge-Kutta Fehlberg
Lion	0.1148731077	0.0001373077
Leopard	0.0022830769	0.0019056154

Table 1 presents a comparison of the Mean Absolute Error (MAE) values for the Euler method and the Runge-Kutta Fehlberg method for these two species. These findings suggest that the Runge-Kutta Fehlberg method is generally more precise than the Euler method.

Figures 5 and 6 show a close-up of the comparison between the exact solution and Euler and RKF methods to compare the accuracy among these two numerical methods. Via the equilibrium and stability tests, this study unveiled four possible cases in the competition dynamics. Case I, species one winning; case II, species two winning; case III, the existence of an unstable equilibrium; and case IV, the coexistence of both species.
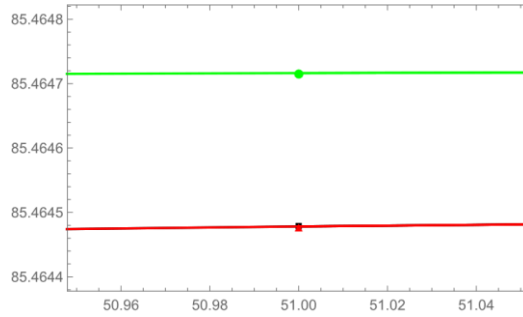
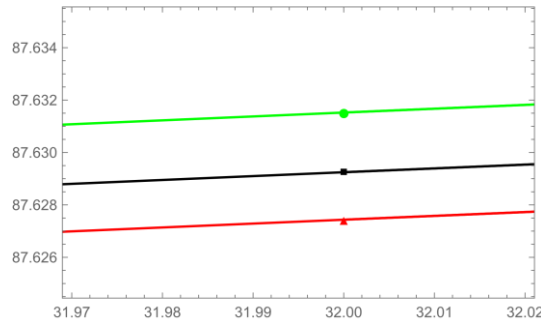Figure 5. Comparison of numerical approximations for lions

Figure 6. Comparison of numerical approximations for leopards

Case I: $\dfrac{K_2}{\beta_{21}} < K_1$ and $\dfrac{K_1}{\beta_{12}} > K_2$, and for Case II is $\dfrac{K_1}{\beta_{12}} < K_2$ and $\dfrac{K_1}{\beta_{12}} < K_2$.

       Figure 7 below shows that the lions' isocline is above and to the right of the leopards. As a result, the leopards are driven to extinction, and the lions' population increase until they reach it carrying capacity (K_1). This stable equilibrium signifies that the lions consistently outcompete the leopards, resulting in the competitive exclusion of the leopards by the lions. While Figure 8 shows that the isocline of leopards is above and to the right of the isocline of lions. Leopards always outcompete lions in this situation, and lions are competitively excluded by leopards.

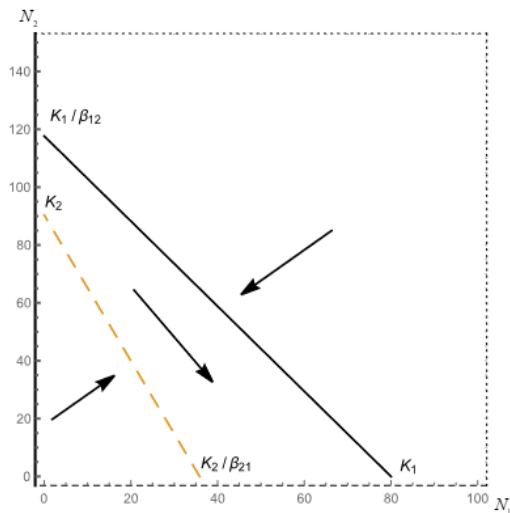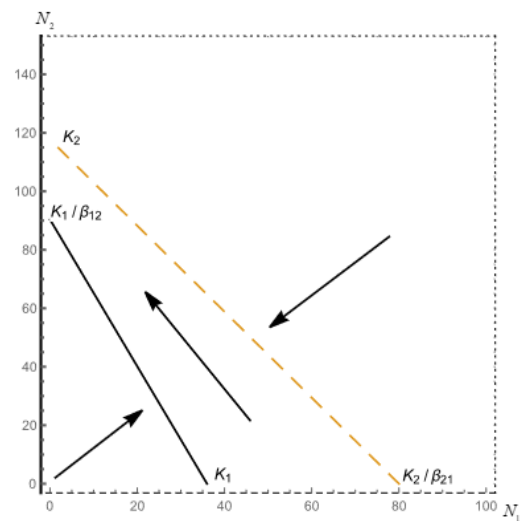Figure 7. Case I (Species 1 wins)

Figure 8. Case II (Species 2 wins)

Case III: $\dfrac{K_2}{\beta_{21}} < K_1$ and $\dfrac{K_1}{\beta_{12}} < K_2$, while Case IV: $\left(\dfrac{K_2}{\beta_{21}}\right) > K_1$ and $\left(\dfrac{K_1}{\beta_{12}}\right) > K_2$.

Figure 9 shows the isoclines of both species cross one another. The carrying capacity of lions (K_1) is higher than the carrying capacity of leopards divided by the competition coefficient $(\dfrac{K_2}{\beta_{21}})$. Besides, the carrying capacity of leopards (K_2) is higher than the carrying capacity of lions divided by the competition coefficient $\left(\dfrac{K_1}{\beta_{12}}\right)$. For Region 2, the point above the dashed line and below the solid line represents the occurrence of competitive exclusion of leopards by lions. While the point above the solid line and below the dashed line in region 1 shows the occurrence of competitive exclusion of lions by leopards.

Figure 10 shows that the isoclines representing the two species, leopards and lions, cross each other. In this case, both species' carrying capacities are lower than the other's carrying capacity divided by the competition coefficient. The species can coexist at a stable equilibrium point, which is the intersection point between both isoclines. At this point, the balance between the two species is achieved because intraspecific competition (competition within the same species) which is stronger than interspecific competition (competition between different species).
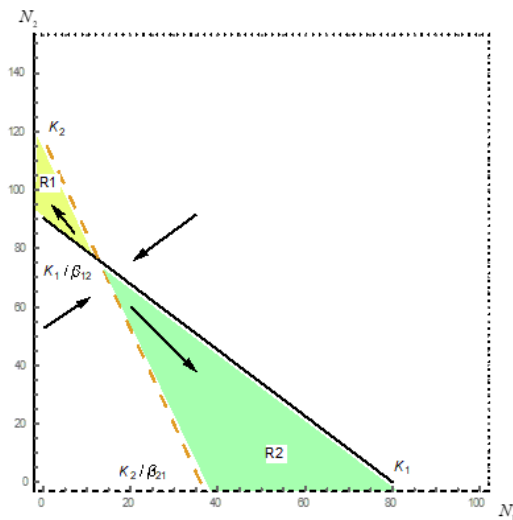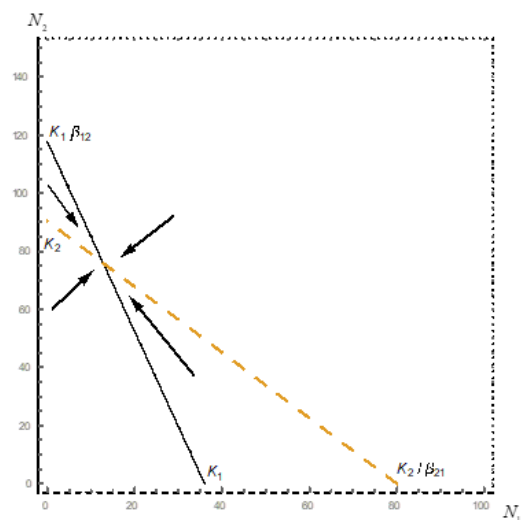

Figure 9. Case III (Unstable equilibrium)   Figure 10. Case IV (Coexistence of both species)

5. Conclusion

In conclusion, this study highlights the efficacy of the Euler and RKF methods for solving the Lotka-Volterra Competitive model. It provides valuable insights into the dynamics of species competition. The results showed that the RKF method provided a more accurate approximation of the numerical methods. The findings also indicated that the study of numerical approaches could serve as reliable prediction tools to eliminate the need for extended observation durations. The estimation of carrying capacity revealed the importance of resource availability in species competition. This suggests that a larger carrying capacity corresponds to a greater ability for a species to thrive and survive in competition. The findings underscored the importance of carrying capacity and initial conditions in understanding equilibrium and stability in competitive interactions. Equilibrium and stability of competition interactions were influenced by the initial conditions and population sizes of the competing species. This study contributes to population dynamics and offers practical implications for wildlife management and conservation efforts. For future research, it is recommended to include more species and consider various other obstacles in species competition. It is suggested to investigate environmental issues, such as the consequences of climate change and how it will affect predator-prey or competitive relationships. Future studies can also examine the performance of numerical approaches while dealing with more complicated interactions and higher-dimensional systems. Aside from that, additional constraints can be included in the model such as spatial barriers or migration patterns. This can provide a more accurate representation of ecological dynamics. Other more sophisticated numerical methods can also be used, such as adaptive step-

size methods, implicit methods, and higher-order Runge-Kutta methods. Moreover, it is fascinating to test solving the model using multiple step methods like Adam-Bashforth, Adam-Moulton and others, and investigate how these methods compare in terms of accuracy and computational efficiency when solving complex ecological models. Other than that, the Lotka-Volterra Competitive model can also be explored for various fields, including business, economic politics, medicine, social sciences and others.

Conflict of Interest
The authors declare there is no conflict of interest in the subject matter or materials discussed in this manuscript.

References

[1]   W. Windarto and E. Eridani, "On modification and application of lotka-volterra competition model," in *AIP Conference Proceedings*, 2020. doi: 10.1063/5.0017242.

[2]   G. F. Gause, "Experimental Studies on the Struggle for Existence," *Journal of Experimental Biology*, vol. 9, no. 4, pp. 389–402, Oct. 1932, doi: 10.1242/jeb.9.4.389.

[3]   C. Marschik, W. Roland, B. Löw-Baselli, and G. Steinbichler, "Application of Hybrid Modeling In Polymer Processing," in *Annual Technical Conference - ANTEC, Conference Proceedings*, 2020.

[4]   S. J. Seytov and D. B. Eshmamatova, "Discrete Dynamical Systems of Lotka–Volterra and Their Applications on the Modeling of the Biogen Cycle in Ecosystem," *Lobachevskii Journal of Mathematics*, vol. 44, no. 4, 2023, doi: 10.1134/S1995080223040248.

[5]   M. Lemos-Silva and D. F. M. Torres, "The Lotka-Volterra Dynamical System and Its Discretization," in *Advanced Mathematical Analysis and its Applications*, 2023. doi: 10.1201/9781003388678-19.

[6]   I. Akjouj *et al.*, "Complex systems in ecology: a guided tour with large Lotka–Volterra models and random matrices," Mar. 06, 2024, *Royal Society Publishing*. doi: 10.1098/rspa.2023.0284.

[7]   N. Ahmad and S. Charan, "Numerical Accuracy Between Runge-Kutta Fehlberg Method and Adams-Bashforth Method For First Order Ordinary Differential Equations With Boundary Value Problem," *J. Math. Comput. Sci*, vol. 6, no. 6, 2016.

[8]   S. Paul, S. P. Mondal, and P. Bhattacharya, "Numerical solution of Lotka Volterra prey predator model by using Runge-Kutta-Fehlberg method and Laplace Adomian decomposition method," *Alexandria Engineering Journal*, vol. 55, no. 1, 2016, doi: 10.1016/j.aej.2015.12.026.

[9]   Muhammad Najmi Abdul Manaf, Nur Fatihah Fauzi, Nurizatul Syarfinas Ahmad Bakhtiar, Nur Izzati Khairudin, and Huda Zuhrah Ab. Halim, "Comparative Analysis of Taylor Series and Runge-Kutta Fehlberg Methods in Solving the Lotka-Volterra Competitive Model," *Applied Mathematics and Computational Intelligence (AMCI)*, vol. 12, no. 3, 2023, doi: 10.58915/amci.v12i3.323.

[10]   B. Huang, Y. Yang, and L. Peng, "Modeling and analysis of ecosystem dynamics based on the Lotka-Volterra model," *Highlights in Science, Engineering and Technology*, vol. 94, pp. 653–660, Apr. 2024, doi: 10.54097/pewswc51.

[11]   M. Clenet, F. Massol, and J. Najim, "Equilibrium and surviving species in a large Lotka-Volterra system of differential equations," May 2022, [Online]. Available: http://arxiv.org/abs/2205.00735

[12]   J. He, Z. Zheng, and Z. Ye, "A new numerical approach method to solve the Lotka–Volterra predator–prey models with discrete delays," *Physica A: Statistical Mechanics and its Applications*, vol. 635, 2024, doi: 10.1016/j.physa.2024.129524.

[13]   N. S. A. B. Razali and F. A. Abdullah, "Numerical methods for Competitive Hunters Model," in *AIP Conference Proceedings*, 2013. doi: 10.1063/1.4801116.

[14]   S. Paul, S. P. Mondal, P. Bhattacharya, and K. Chaudhuri, "Some comparison of solutions by different numerical techniques on mathematical biology problem," *International Journal of Differential Equations*, vol. 2016, 2016, doi: 10.1155/2016/8921710.

[15]   T. D. Chandra, M. Yasin, and F. Lailiyah, "Numerical solution to the differential equation system of Lotka-Volterra by using Heun method," in *AIP Conference Proceedings*, 2024. doi: 10.1063/5.0194394.

[16]   S. Han, "Plant Community Changes Based on The Lotka-Volterra Model," *Highlights in Science, Engineering and Technology*, vol. 63, 2023, doi: 10.54097/hset.v63i.10856.

[17]   J. Wang, Q. P. Liu, and Y. S. Luo, "The Numerical Analysis of the Long Time Asymptotic Behavior for Lotka-Volterra Competition Model with Diffusion," *Numer Funct Anal Optim*, vol. 40, no. 6, 2019, doi: 10.1080/01630563.2019.1566245.

[18]   S. Chatterjee and M. Acharyya, "Critical slowing down along the separatrix of Lotka-Volterra model of competition," *International Journal of Modern Physics C*, vol. 34, no. 9, 2023, doi: 10.1142/S0129183123501188.

[19]   G. A. Balme, R. T. Pitman, H. S. Robinson, J. R. B. Miller, P. J. Funston, and L. T. B. Hunter, "Leopard distribution and abundance is unaffected by interference competition with lions," *Behavioral Ecology*, vol. 28, no. 5, 2017, doi: 10.1093/beheco/arx098.

Biography of all authors

Picture	Biography	Authorship contribution
	Nur Hidayah Abd Rahaman with Bachelor of Science (Hons.) Management Mathematics from College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Perlis.	Design the research work and data collection.
	Nurizatul Syarfinas Ahmad Bakhtiar obtained her first degree in Pure Mathematics from the University Sains Malaysia, Malaysia in 2011. She received a master's and PhD degree in Mathematical Modeling from the same university in 2012 and 2020 respectively. Currently, she is a senior lecturer at the Universiti Teknologi MARA Perlis. Her research interests include dissipative solitons and dynamical systems and analysis.	Design the research work and drafting article.
	Hafizah Hajimia obtained herfirst degree in TESL at Universiti Pendidikan Sultan Idris. She received a master's and PhD from Universiti Utara Malaysia. Currently, she is a senior lecturer at the Universiti Utara Malaysia.	Design the research work.
	Nur Fatihah Fauzi obtained her first degree in Pure Mathematics from the University Sains Malaysia, Malaysia in 2011. She received a master's and PhD degree in Mathematical Modeling from the same university in 2012 and 2018 respectively. Currently, she is a senior lecturer at the Universiti Teknologi MARA Perlis. Her research interests include boundary layer flow and heat transfer.	Data collection, data analysis and interpretation.
	Nur Izzati Khairudin obtained her first degree in Pure Mathematics from the University Sains Malaysia, Malaysia in 2011. She received a master's and PhD degree in Mathematical Modeling from the same university in 2012 and 2017 respectively. Currently, she is a senior lecturer at the Universiti Teknologi MARA Perlis. Her research interests include dissipative solitons and dynamical systems and analysis.	Analysis of data and interpretation.

Assessing Flood Risk Using L-Moments: An Analysis of the Generalized Logistic Distribution and the Generalized Extreme Value Distribution at Sayong River Station

Nur Diana Zamani
Kolej Pengajian Perkomputeran, Informatik dan Matematik, Universiti Teknologi MARA Cawangan Johor Kampus Segamat, Malaysia
nurdi958@uitm.edu.my

Basri Badyalina
Kolej Pengajian Perkomputeran, Informatik dan Matematik, Universiti Teknologi MARA Cawangan Johor Kampus Segamat, Malaysia
basribdy@uitm.edu.my

Muhammad Zulqarnain Hakim Bin Abd Jalal
Kolej Pengajian Perkomputeran, Informatik dan Matematik, Universiti Teknologi MARA Cawangan Johor Kampus Segamat, Malaysia
zulqarnainhakim@uitm.edu.my

Rusnani Mohamad Khalid
Kolej Pengajian Perkomputeran, Informatik dan Matematik, Universiti Teknologi MARA Cawangan Johor Kampus Segamat, Malaysia
rusna162@uitm.edu.my

Fatin Farazh Ya'acob
Fakulti Pengurusan dan Perniagaan, Universiti Teknologi MARA Cawangan Johor Kampus Segamat, Malaysia
fatinfarazh@uitm.edu.my

Kerk Lee Chang
Kolej Pengajian Perkomputeran, Informatik dan Matematik, Universiti Teknologi MARA Cawangan Johor Kampus Segamat, Malaysia
kerkleechang@uitm.edu.my

ABSTRACT

This study examines severe flood events at Sayong River Station by conducting a Flood Frequency Analysis using the Generalized Logistic (GLO) and Generalized Extreme Value (GEV) distributions. The L-moment approach is utilized for parameter estimation, with quantile estimates assessed for return periods of 10, 50, and 100 years. A comprehensive comparison of statistical performance indicators, such as RMSE, MAE, and MAPE, was performed to identify the best realistic model for depicting severe flood behavior. The findings indicate that the GLO distribution consistently outperforms the GEV distribution in all criteria. The GLO distribution demonstrated superior performance with a lower RMSE (17.7369), MAE (8.6608), and MAPE (11.83%) relative to the GEV distribution, which exhibited an RMSE of 17.8034, MAE of 8.7957, and MAPE of 12.98%. These findings validate the GLO distribution as the better appropriate model for representing peak streamflow data. Moreover, quantile estimates obtained from the GLO distribution are 197.3153 m³/s for the 10-year, 363.8308 m³/s for the 50-year and 469.9711 m³/s for the 100-year return periods. The GLO distribution exhibit greater concordance with empirical data, further validating its accuracy. The superior performance of the GLO distribution emphasizes the importance of selecting the appropriate distribution for flood risk assessment. The GLO distribution yields more accurate predictions of severe flood magnitudes, hence enhancing flood estimations, infrastructure design, and mitigation measures at Sayong River Station.

PENERBIT *PRESS*
UNIVERSITI TEKNOLOGI MARA

Corresponding Author:
Basri Badyalina
Universiti Teknologi MARA Cawangan Johor Kampus Segamat, Malaysia
Email: basribdy@uitm.edu.my

1. Introduction

Natural disasters like floods are very destructive in Malaysia, causing damage on lives of individuals, agricultural products, and infrastructure [1-3]. The nature of environment in Malaysia characterized by monsoonal rainfall and frequent thunderstorms, makes it particularly vulnerable to flash floods [4]. During periods of intense precipitation or severe weather phenomena, the river is prone to flooding, which can affect the community, infrastructure, and the surrounding environment. Given the circumstances, there is an urgent need for accurate assessment and water resource management.

Flood Frequency Analysis (FFA) is an essential tool can be used to the estimate the probability and scale of future flood occurrences [5]. FFA provides essential insights into the likelihood and severity of flooding events. By estimating the frequency and magnitude of extreme floods, stakeholders can better understand the risks associated with flood hazards, enabling them to implement appropriate mitigation measures. As climate change continues to influence weather patterns and hydrological systems, FFA can provide insights into changing flood risks [6]. This helps communities adapt to new flood dynamics, informing strategies for resilience in the face of climate-related challenges.

Currently, there is ongoing interest in determining the optimal distribution for FFA. Researchers are exploring various statistical models, such as the Gumbel, Log-Pearson Type III, and Generalized Extreme Value distributions, to identify which model best fits historical flood data. A significant method gaining traction in this context is the use of L-Moments, which offer a robust alternative to traditional moment-based approaches. Moreover, the integration of L-Moments into the FFA process allows for the calculation of key statistical measures, such as the L-CV (L-Coefficient of Variation) and L-Skewness. These measures provide insights into the distribution of flood data, enhancing the understanding of flood risk and facilitating more accurate flood modelling.

By applying these methodologies, researchers can develop a more nuanced understanding of the underlying hydrological processes that contribute to flooding. Current research efforts are focused on determining the optimal statistical distributions for FFA, which can significantly enhance the accuracy of flood predictions. By refining these methodologies, researchers aim to improve the reliability of flood frequency estimates, enabling more effective disaster preparedness and response initiatives.

2. Literature Review

Through the use of historical flood data, FFA allows researchers and policymakers to evaluate the probability of flood events occurring within a specified time frame, known as return periods at targeted location [7]. Precise estimations of these return period are essential for effective flood- mitigating planning and strategic emergency responses management. FFA is particularly crucial in Malaysia, given the potential for floods to inflict significant damage on both rural and urban areas [8]. The necessity to adjust to the changing precipitation pattern and flood behaviors driven by climate change further highlights the significance of FFA.

This paper aims to conduct FFA for the Sayong River in Johor, Malaysia, utilizing two most implemented distribution in Malaysia namely GEV and GLO distributions for peak streamflow data [9],[10]. This study presents a thorough comparison between the Generalized Extreme Value (GEV) and Generalized Logistic (GLO) distributions in modelling flood events in Malaysia. The analysis reveals the relative strengths and weaknesses of each distribution, providing valuable insights into their applicability in different hydrological contexts. FFA plays a critical role in assessing flood risks, especially in regions like Malaysia, where extreme hydrological events are frequent.

The GEV and GLO distributions are often used in FFA because of their adaptability and resilience in representing extreme hydrological occurrences [10-12]. The GEV distribution comprising Gumbel, Fréchet, and Weibull distributions. The GEV and GLO distributions are often used statistical distributions in FFA because of their adaptability and resilience in representing extreme hydrological occurrences [11-13]. The GEV distribution comprises three unique statistical distributions: Gumbel,

Fréchet, and Weibull distributions [14]. The GEV distribution offers flexibility in handling hydrological data and is valuable for evaluating extreme event likelihood by considering block maxima in datasets [15]. It is particularly suitable for analyzing yearly maximum flood levels and has been proven suitable in FFA. The GEV distribution is specifically designed to model the distribution of extreme values, making it ideal for analyzing peak flows, which are critical for flood risk assessment and management. It is well-suited for analyzing yearly maximum flood levels and has been proven suitable in FFA.

The GLO distribution recognized for its logistic-type tail characteristics, provides an alternative approach for modelling extreme events, especially when the data exhibits asymmetry or heavy tails [16]. Previous studies in Peninsular Malaysia have utilized various statistical distributions to measure flood frequency, with GEV and GLO distributions being the most used options for FFA in Malaysia. Ahmad *et al.* [9] finds that the Generalized Extreme Value (GEV) distribution model is the best-fitted flood distribution model for the Kuantan River Basin, with a P-value of 0.997 compare to other distributions such as Generalized Pareto distribution, Log Pearson (3) distribution, Weibull distribution and Log-normal distribution.

Badyalina *et al.* [17] implement GEV and GLO distribution to fit to the annual peak streamflow for the Segamat river. In addition, the parameters of GEV and GLO distribution are estimated using the L-Moment method. The result shows that the GEV and GLO distribution well fitted the annual peak streamflow of the Segamat River. Ismail *et al.* [18] examines and compares five flood distribution models for the Johor River basin in Malaysia. The investigation suggests that the GEV model is the most suitable for representing the yearly peak flow data. The GEV model outperformed other distributions such as Pearson 3, Lognormal, Weibull, and Gamma.

The GLO distribution is particularly effective for modelling extreme flood events, which are often characterized by high variability and non-normality [19]. Its ability to capture the tail behavior of the data is crucial for accurate flood frequency analysis. Malaysia experiences a tropical climate characterized by intense and irregular rainfall patterns, leading to sudden and extreme flood events. The GEV and GLO distribution can effectively capture the behavior of such extreme hydrological data.

A key aspect of FFA is the precise estimation of the parameters of statistical distributions used to represent annual peak flow data. The L-moments method is a reliable statistical analytical technique that is increasingly employed in hydrology for this task [2], [11], [20-22]. Precise parameter estimation using L-moments requires solving a set of nonlinear equations, which presents some challenges. The L-moments method is less affected by data length and hence produces more precise results in comparison to the method of moment technique [23].

The benefits of the L-moment technique used in this study are as follows: wide applicability, robustness to outliers and minimal bias [24], [25]. The attributes of L-Moment method make them particularly suitable for regions like Malaysia, where flood events can vary significantly and are influenced by extreme weather. L-moments offer enhanced reliability and stability in estimating distribution parameters, particularly when managing extremely high values, making them crucial for precise measurement of FFA. L-moments can produce stable estimates even with small sample sizes, which is crucial in hydrology, where data for extreme events may be limited [26].

The objective of this study is to apply FFA to Saying River to determine which distribution best fits the annual peak flow data. The selected distribution will then be used to estimate the return period, providing insights into the expected frequency and severity of future floods for policymakers, thereby improving the effectiveness of water management and flood risk assessment.

3. Methodology

3.1 Data

The station selected for this study is the Sungai Sayong River Station, located in Johor, Malaysia. The data is obtained from Department of Irrigation and Drainage, Malaysia. Peak flow data helps identify and quantify extreme flow events, which are critical for assessing flood risk in a specific area. Understanding the frequency and magnitude of peak flows aids in preparing for potential flooding scenarios. Table 1 presents the descriptive statistics of the streamflow characteristics for Sayong River, including the mean, median, standard deviation, variance, kurtosis, and skewness. The dataset was collected over a period of 33 years, starting from 1984.

Table 1. Descriptive Statistics for Sayong River Streamflow					
Mean	Median	Standard Deviation	Variance	Kurtosis	Skewness
113.4379	87.99	85.79689	7361.106	9.677445	2.714384

The descriptive statistics for streamflow at the Sayong River station are shown in Table 1. The average flow rate was 113.44 m³/s. The median streamflow was 87.99 m³/s, which is significantly lower than the mean, suggesting the data is highly skewed to the right. A high standard deviation of 85.80 m³/s and indicate considerable variation in streamflow which means there exist extreme value in the data. The positive skewness value of 2.71 indicates that the data distribution is right-skewed. Additionally, a kurtosis score of 9.68 indicates a distribution with heavy tails and a higher frequency of extreme values. Based on the descriptive statistics presented in Table 1, the streamflow at Sayong River station shows evidence of extreme value, suggesting the occurrence of extreme events which are floods.

3.2 GEV Distribution

The Generalized Extreme Value (GEV) distribution is a flexible and extensively employed statistical distribution in hydrology, specifically for representing extreme events like floods or phenomena where only the most extreme values are of significance. The GEV distribution unifies three types of distributions - Gumbel, Fréchet, and Weibull into a single family, distinguished by its adaptability in modelling various extreme events. Its versatility in accommodating various data formats and its capacity to represent both heavy-tailed and bounded distributions make it highly suitable like Malaysia, where the characteristics of flood events can vary greatly among different river basins. The cumulative density function (CDF) of GEV is defined in Eq. 1 and Eq. 2:

$$f(x) = \frac{1}{\alpha}\left[1-k\left(\frac{x-\xi}{\alpha}\right)\right]^{\frac{1}{k}-1} \exp\left\{-\left[1-k\left(\frac{x-\xi}{\alpha}\right)\right]^{\frac{1}{k}}\right\} \tag{1}$$

The quantile function of GEV distribution can be defined as

$$x(F) = \hat{\xi} + \frac{\hat{\alpha}}{\hat{k}}\left\{1-(-\ln(F))^{\hat{k}}\right\} \tag{2}$$

The quantile function is essentially the inverse of the CDF (Cumulative Density Function). While the CDF gives the probability of a variable being below a certain threshold x, the quantile function (also called the percent-point function) tells us what value of x corresponds to a specific cumulative probability. The definition to estimate the parameter $\hat{\xi}, \hat{\alpha}, \hat{k}$ definition from Hosking et $al.$ [27]. The parameters $\hat{\xi}, \hat{\alpha}$ and $\hat{k}$ are estimated according to the definitions provided by Hosking et $al.$ [27]. The definition of $\hat{\xi}, \hat{\alpha}$ and $\hat{k}$ for GEV distribution are provided in Eq. 3 to Eq. 5:

$$\hat{k} = 7.85890c + 2.9554c^2 \tag{3}$$

where $c = \frac{2}{3+t_3} - \frac{\ln 2}{\ln 3}$

$$\hat{\alpha} = \frac{l_2}{\Gamma(\hat{k})(1-2^{\hat{k}})} \tag{4}$$

$$\hat{\xi} = l_1 - \frac{\hat{\alpha}}{\hat{k}} + \hat{\alpha}\Gamma(\hat{k}) \tag{5}$$

3.3 GLO Distribution

Generalized Logistic (GLO) distribution is used to model the distribution of extreme values, such as high annual maximum flood flows. It is especially effective in scenarios where the tails of the distribution (representing rare and extreme events) need to be accurately represented. This distribution can help predict the likelihood of extreme floods based on historical data.

The probability density function (PDF) of GLO distribution is given by

$$f(x) = \frac{1}{\alpha}\left\{1-k\left(\frac{x-\xi}{\alpha}\right)\right\}^{\frac{1}{k}-1}\left[1+\left\{1-k\left(\frac{x-\xi}{\alpha}\right)\right\}^{\frac{1}{k}}\right]^{-2} \tag{6}$$

The quantile function of GLO distribution can be written as:

$$x(F) = \hat{\xi} + \frac{\hat{\alpha}}{\hat{k}}\left[1-\left\{\frac{(1-F)}{F}\right\}^{\hat{k}}\right] \tag{7}$$

The definition to estimate the parameter $\hat{\xi}, \hat{\alpha}, \hat{k}$ definition from Hosking *et al.* [27]. The parameters $\hat{\xi}, \hat{\alpha}$ and $\hat{k}$ are estimated according to the definitions provided by Hosking *et al.* [27]. The definition of $\hat{\xi}, \hat{\alpha}$ and $\hat{k}$ for GLO distribution are provided in Eq. 8 to Eq. 10:

$$\hat{k} = -t_3 \tag{8}$$

$$\hat{\alpha} = \frac{l_2}{\Gamma(\hat{k})\left[\Gamma(1-\hat{k})-\Gamma(2-\hat{k})\right]} \tag{9}$$

$$\hat{\varepsilon} = l_1 - \frac{\hat{\alpha}}{\hat{k}} + \hat{\alpha}\Gamma(\hat{k})\Gamma(1-\hat{k}) \tag{10}$$

3.4 Parameter estimation using L-Moment

The L-moment methodology is a statistical technique used for estimating the parameters of probability distributions, particularly in hydrology and environmental sciences. L-moments are analogous to conventional moments (like mean and variance) but offer some advantages, especially when dealing with skewed data or outliers. L-moments are linear combinations of order statistics, which are the sorted values of a sample. Unlike conventional moments, L-moments are less influenced by extreme values or outliers, making them more robust for skewed data. The unbiased sample estimator of the L-Moment method defined by Landwehr *et al.* [28] is:

$$b_r = \frac{1}{n}\binom{n-1}{r}^{-1}\sum_{i=r+1}^{n}\binom{i-1}{r}x_{i:n} \quad r = 0, 1, 2, \ldots \tag{11}$$

The first four L-moments sample estimates can be written as:

$$l_1 = b_0 \tag{12}$$
$$l_2 = 2b_1 - b_0 \tag{13}$$
$$l_3 = 6b_2 - 6b_1 + b_0 \tag{14}$$
$$l_4 = 20b_3 - 30b_2 + 12b_1 - b_0 \tag{15}$$

The L-moments ratio samples are defined as:

$$t_2 = \frac{l_2}{l_1} \tag{16}$$

$$t_3 = \frac{l_3}{l_2} \tag{17}$$

$$t_4 = \frac{l_4}{l_2} \tag{18}$$

t_2 refer to coefficient of variations, t_3 refer to coefficient of skewness and t_4 refer to kurtosis.

3.5 Prediction Error Metrics

An essential component of FFA is choosing a suitable statistical distribution that accurately represents the observed data. Each distribution provides distinct characteristics extreme events and selecting the most appropriate one can greatly influence the reliability of the estimated flood event. In this study three different metrics implemented to determine the best fitted candidate distribution (GEV and GLO) distribution. Most used metrics for evaluating model performance are Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). The definition for MAPE, MAE, and RMSE is given in the series of equations below Eq.19 to Eq.21, respectively.

$$MAPE = \left(\frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \right) \times 100\% \tag{19}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2} \tag{20}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - \hat{y}_i \right| \tag{21}$$

where y_i is the observed flows, $\hat{y}_i$ is the predicted flows and n is the number of flow series that have been model.

4. Results and Discussion

FFA is a fundamental tool in hydrological, providing insights into the magnitude and frequency of flood events. These analyses are crucial for guiding water resource management and flood mitigation strategies. In this study, the yearly maximum streamflow data, spanning 33 years, were fitted to the GEV and GLO distributions for modelling extreme flood. This data allows for accurate flood frequency analysis and helps design resilient infrastructure. The parameters of GLO and GEV are estimated using the L-Moment method. The parameters of both distributions were estimated using the L-Moment method. The process involves calculating the first four unbiased sample estimators using Eq.11. The value of first four of unbiased sample estimator are:

$$b_0 = 111.4379$$
$$b_1 = 76.5969$$
$$b_2 = 60.1362$$
$$b_3 = 50.5743$$

Once the first four unbiased sample estimators have been obtained, the first four L-moment sample estimates can be calculated by using Eq. 12 to Eq. 15. The values for the first four L-Moment are:

$$l_1 = 113.4379$$
$$l_2 = 39.7559$$
$$l_3 = 14.6739$$
$$l_4 = 13.1242$$

Using Eq.16 to Eq. 18, L-Moment ratios (t_2, t_3, t_4) can be obtained. The values for the L-moment ratios are:

$$t_2 = 0.3505$$
$$t_3 = 0.3691$$
$$t_4 = 0.3301$$

Using the L-Moment ratios values, by implementing a methodology by Hosking *et al.* [27], the parameter for GEV and GLO can be estimated. The estimated parameters for GEV and GLO distribution.

Table 2. Estimated Parameters using L-Moments for GEV and GLO

Distribution	Parameters		
GEV	$k = -0.289$	$u = 74.121$	$\alpha = 40.439$
GLO	$k = -0.37$	$\varepsilon = 90.872$	$\alpha = 31.427$

Table 2 shown the estimated parameter for the GEV and GLO distributions using L-Moment. The parameters for GEV distribution are $k = -0.289$, $u = 74.121$ and $\alpha = 40.439$. The parameters value for GLO distribution are $k = -0.37$, $\varepsilon = 90.872$ and $\alpha = 31.427$. To determine the most suitable candidate distribution for extreme flood events at Sayong River station, three different metrices are used namely RMSE, MAE and MAPE. Evaluating these measures enables the identification of the distribution that best represents the extreme flood behaviour. A summary of the RMSE, MAE, and MAPE results for the data from Sayong River is presented in Table 3.

Table 3. RMSE, MAE and MAPE at Sayong River Station

Distribution	RMSE		MAE		MAPE	
	Value	Rank	Value	Rank	Value	Rank
GLO	17.7369	1	8.6608	1	11.83%	1
GEV	17.8034	2	8.7957	2	12.98%	2

A thorough assessment of the GLO and GEV distributions, as shown in Table 3, indicates that the GLO distribution offers a superior fit for extreme flood events at Sayong River Station. The GLO distribution exhibits a lower RMSE of 17.7369 in contrast to the GEV 17.8034, indicating that the GLO distribution often yields smaller prediction errors. In terms of MAE value, the GLO distribution also produces lower MAE value which is 8.6608 meanwhile for GEV distribution is 8.7957. The MAPE value for GLO distribution is 11.83% lower than MAPE value for GEV value which is 12.98%. These results indicate that the GLO distribution performs better than the GEV distribution in all assessed metrics. Thus, GLO distribution is selected to represent Sayong river peak streamflow data.

The GLO distribution will be used to estimate the flood occurrence at Sayong river station. A precise estimation of quantiles from the GLO distribution is essential in flood event analysis, as it provides valuable insights into the likelihood of various streamflow levels based on their return periods. Quantile estimates for extreme flood events at Sayong River Station, computed for return periods of 10, 50, and 100 years. The streamflow level is expected to be exceeded once every ten years, as indicated by the 10-year return period projection. This is crucial for strategic long-term planning and infrastructure development, as the 50-year return period suggests a more intense flood event with a 2% likelihood of surpassing projected levels annually. Critical safety planning and design are conducted using the 100-year return period, which is equivalent to a flood occurrence with a 1% annual probability. The return periods for 10, 50, and 100 years are Shown in Table 4.

Table 4. Estimation of T=10,50 and 100 years			
T	10	50	100
Q_T	197.3153 m³/s	363.8308 m³/s	469.9711 m³/s

Table 4 presents the quantile estimated for the Sayong River Station for return periods of 10, 50, and 100 years. The 10-year return period estimate of 197.3153 m³/s indicate that the streamflow level is projected to be surpassed on average once every ten years. This value denotes the extent of a very common yet substantial flood occurrence. For the 50-year return period, the calculated quantile of 363.8308 m³/s suggests a more severe flood occurrence with a 2% probability of being surpassed in any given year. An estimated for 100-year return period quantile of 469.9711 m³/s corresponds to a streamflow level that has a 1% probability of being exceeded in any given year. These return period information at Sayong River is crucial for efficient water management and flood mitigation projects.
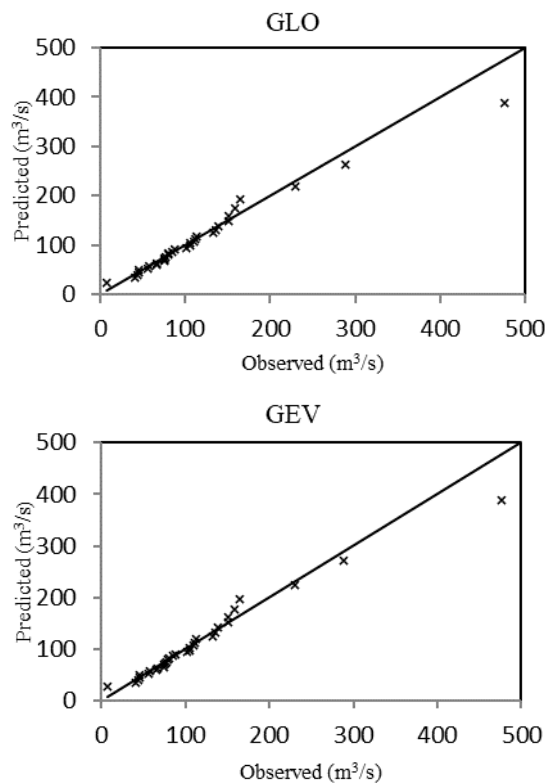

Figure 1. Observed vs. Estimated flood peak value at Sayong river station for GEV and GLO distribution.

5. Conclusion

This study implements FFA to determine the flood occurrences at Sayong River Station through the implementation of GEV and GLO distributions. L moments are used to estimate the parameter of distribution. Then, the parameters for GEV and GLO distributions are utilized to estimate flood quantiles with various return periods. The return periods selected in this study are 10, 50 and 100 years. The results indicate that the GLO distribution fits the peak streamflow data in comparison to the GEV distribution in terms of three different metrices (RMSE, MAE and MAPE). Specifically, the quantile estimates obtained for the 10-year, 50-year, and 100-year return periods were 197.3153 m³/s, 363.8308 m³/s, and 469.9711 m³/s, respectively. Policymakers can use the quantile estimate result from this study for water resource planning at the Sayong River station.

Conflict of Interest

The authors declare there is no conflict of interest in the subject matter or materials discussed in this manuscript.

References

[1]   K. J. Ghalehteimouri, F. C. Ros, and S. Rambat, "Flood risk assessment through rapid urbanization LULC change with destruction of urban green infrastructures based on NASA Landsat time series data: A case of study Kuala Lumpur between 1990–2021," *Ecological Frontiers,* vol. 44, no. 2, pp. 289-306, 2024, doi: 10.1016/j.chnaes.2023.06.007.

[2]   N. H. Hassim *et al.*, "In Situ Flood Frequency Analysis Used for Water Resource Management in Kelantan River Basin," *AL JOUR,* vol. 11, no. 5, p. 9, 2022.

[3]   N. A. M. Jan, A. Shabri, and B. Badyalina, "Selecting probability distribution for regions of Peninsular Malaysia streamflow," in *AIP Conference Proceedings*, 2016, vol. 1750, no. 1: AIP Publishing, doi: 10.1063/1.4954619.

[4]   W. H. M. W. Mohtar, J. Abdullah, K. N. A. Maulud, and N. S. Muhammad, "Urban flash flood index based on historical rainfall events," *Sustainable Cities and Society,* vol. 56, p. 102088, 2020, doi: 10.1016/j.scs.2020.102088.

[5]   G. Perez, E. T. Coon, S. S. Rathore, and P. V. Le, "Advancing process-based flood frequency analysis for assessing flood hazard and population flood exposure," *Journal of Hydrology,* vol. 639, p. 131620, 2024, doi: 10.1016/j.jhydrol.2024.131620.

[6]   C. Wasko *et al.*, "Incorporating climate change in flood estimation guidance," *Philosophical Transactions of the Royal Society A,* vol. 379, no. 2195, p. 20190548, 2021, doi: 10.1098/rsta.2019.0548.

[7]   M. Islam *et al.*, "Discover Applied Sciences," *Discover,* vol. 6, p. 174, 2024.

[8]   S. A. A. Talib, W. M. R. Idris, L. J. Neng, T. Lihan, and M. Z. A. Rasid, "Irregularity and time series trend analysis of rainfall in Johor, Malaysia," *Heliyon,* vol. 10, no. 9, 2024, doi: 10.2139/ssrn.4435183.

[9]   A. M. Ahmad, N. S. Romali, and S. Sulong, "Flood frequency analysis of annual maximum stream flows for Kuantan River Basin," in *AIP Conference Proceedings*, 2023, vol. 2688, no. 1: AIP Publishing, doi: 10.1063/5.0111746.

[10]   N. A. M. Jan, A. Shabri, M. F. Marsani, and B. Badyalina, "TL-Moments Approach: Application of non-Stationary GEV Model in Flood Frequency Analysis," 2021, doi: 10.21203/rs.3.rs-673867/v1.

[11]   C. T. Vidrio-Sahagún, J. Ruschkowski, J. He, and A. Pietroniro, "A practice-oriented framework for stationary and nonstationary flood frequency analysis," *Environmental Modelling & Software,* vol. 173, p. 105940, 2024, doi: 10.1016/j.envsoft.2024.105940.

[12]   A. Ahmed, Z. Khan, and A. Rahman, "Searching for homogeneous regions in regional flood frequency analysis for Southeast Australia," *Journal of Hydrology: Regional Studies,* vol. 53, p. 101782, 2024, doi: 10.1016/j.ejrh.2024.101782.

[13]   N. A. M. Jan, A. Shabri, J. Hounkpè, and B. Badyalina, "Modelling non-stationary extreme streamflow in Peninsular Malaysia," *International Journal of Water,* vol. 12, no. 2, pp. 116-140, 2018, doi: 10.1504/ijw.2018.091380.

[14]   M. M. El Genidy, E. A. Hebeshy, B. E. S. El Desouky, and R. S. Gomaa, "An Accurate Method for Estimating the Parameters of the Generalized Extreme Value Distribution Using its Moments," *Alfarama Journal of Basic & Applied Sciences,* vol. 5, no. 2, pp. 190-207, 2024, doi: 10.21608/ajbas.2023.236538.1183.

[15]   N. K. Goel and V. Rajendran, "Modelling climate change-induced nonstationarity in rainfall extremes: A comprehensive approach for hydrological analysis," *Technological Forecasting and Social Change,* vol. 208, p. 123693, 2024, doi: 10.1016/j.techfore.2024.123693.

[16]   A.-E. A. Teamah, A. A. Elbanna, and A. M. Gemeay, "Heavy-tailed log-logistic distribution: Properties, risk measures and applications," *Statistics, Optimization & Information Computing,* vol. 9, no. 4, pp. 910-941, 2021, doi: 10.19139/soic-2310-5070-1220.

[17]   B. Badyalina, N. A. Mokhtar, N. A. M. Jan, N. H. Hassim, and H. Yusop, "Flood frequency analysis using L-moment for Segamat river," *Matematika,* pp. 47-62, 2021, doi: 10.11113/matematika.v37.n2.1332.

[18]   A. Z. Ismail, Z. Yusop, and Z. Yusof, "Comparison of flood distribution models for Johor River basin," *Jurnal Teknologi,* vol. 74, no. 11, 2015, doi: 10.11113/jt.v74.4881.

[19]   Z. Zhang, T. A. Stadnyk, and D. H. Burn, "Identification of a preferred statistical distribution for at-site flood frequency analysis in Canada," *Canadian Water Resources Journal/Revue canadienne des ressources hydriques,* vol. 45, no. 1, pp. 43-58, 2020, doi: 10.1080/07011784.2019.1691942.

[20]   C. Anghel, S. C. Stanca, and C. Ilinca, "Exploring the Applicability and Insights of the Pearson Type III Distribution in Flood Frequency Analysis," *Revista Romana de Inginerie Civila,* vol. 15, no. 3, pp. 1-13, 2024, doi: 10.37789/rjce.2024.15.3.6.

[21]   M. F. Marsani, A. Shabri, B. Badyalina, N. A. M. Jan, and M. S. M. Kasihmuddin, "Efficient market hypothesis for Malaysian extreme stock return: Peaks over a threshold method," *Matematika,* pp. 141-155, 2022, doi: 10.11113/matematika.v38.n2.1396.

[22]   N. A. M. Jan, A. Shabri, S. Ismail, B. Badyalina, S. S. Abadan, and N. Yusof, "Three-parameter lognormal distribution: Parametric estimation using L-moment and TL-moment approach," *Jurnal Teknologi,* vol. 78, no. 6-11, 2016, doi: 10.11113/jt.v78.9202.

[23]   C. Ilinca and C. G. Anghel, "Flood frequency analysis using the gamma family probability distributions," *Water,* vol. 15, no. 7, p. 1389, 2023, doi: 10.20944/preprints202303.0320.v1.

[24]   Y. Liang, S. Liu, Y. Guo, and H. Hua, "L-moment-based regional frequency analysis of annual extreme precipitation and its uncertainty analysis," *Water Resources Management,* vol. 31, pp. 3899-3919, 2017, doi: 10.1007/s11269-017-1715-5.

[25]   R. M. Vogel and N. M. Fennessey, "L moment diagrams should replace product moment diagrams," *Water resources research,* vol. 29, no. 6, pp. 1745-1752, 1993, doi: 10.1029/93wr00341.

[26]   I. Dekker, K. Dubrawski, P. Jones, and R. MacDonald, "Estimating Non-Stationary Extreme-Value Probability Distribution Shifts and Their Parameters Under Climate Change Using L-Moments and L-Moment Ratio Diagrams: A Case Study of Hydrologic Drought in the Goat River Near Creston, British Columbia, ". *Hydrology*, vol. 11, no. 9, p. 154, 2024, doi: 10.3390/hydrology11090154.

[27]   J. R. M. Hosking, J. R. Wallis, and E. F. Wood, "Estimation of the generalized extreme-value distribution by the method of probability-weighted moments," *Technometrics,* vol. 27, no. 3, pp. 251-261, 1985, doi: 10.21236/ada141696.

[28]   J. M. Landwehr, N. Matalas, and J. Wallis, "Probability weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles," *Water resources research,* vol. 15, no. 5, pp. 1055-1064, 1979, doi: 10.1029/wr015i005p01055.

Biography of all authors

Picture	Biography	Authorship contribution
	Nur Diana Zamani is a senior lecturer from the Universiti Teknologi MARA Johor branch.	Design the research work, interpretation, and draft article.
	Basri Badyalina is a senior lecturer from the Universiti Teknologi MARA Johor branch.	Development of modelling and programming.
	Muhammad Zulqarnain Hakim Bin Abd Jalal is a senior lecturer from the Universiti Teknologi MARA Johor branch.	Development of literature review
	Rusnani Mohamad Khalid is a senior lecturer from the Universiti Teknologi MARA Johor branch.	Development of modelling and programming.
	Fatin Farazh Ya'acob is a senior lecturer from the Universiti Teknologi MARA Johor branch.	Draft article.
	Kerk Lee Chang is a senior lecturer from the Universiti Teknologi MARA Johor branch.	Collection of environmental data.