Menstruation Cycle Information Analysis for Pattern Recognition: Determination of Algorithm on Stakeholder Requirement

ED Uthpala[1#], ERC Sandamali[2] and RMM Pradeep[1]

[1]*Department of Information Technology, Faculty of Computing, General Sir John Kotelawala Defence University, Sri Lanka*

[2]*Department of Computational Mathematics, Faculty of Computing, General Sir John Kotelawala Defence University, Sri Lanka*

[#]35-it-0036@kdu.ac.lk

Abstract - The menstrual cycle of a healthy woman is systematic and individually unique. As it is directly related to a woman's physical and mental health, the menstrual cycle plays a major role in individual nutrition, social and psychological decision-making. As women frequently forget the exact date of menstruation, lots of mobile apps are developed to assist them. All such apps use 28 as the approximate date, but the experiences are very dependent. Therefore, to utilize individualized menstruation cycle guidance app development, it is required to develop an algorithm to predict the date of menstruation. Then the objective of this work is to study the collection and analysis of field data to realize what model is suited for cycle prediction. The data was collected using 30 women between the ages of 20-35 with their menstrual cycle dates for one year. Then this time series data was analysed using cumulative moving average (CMA), and Auto-Regressive Integrated Moving Average (ARIMA). The analysis shows both methods can predict menstrual dates with an average accuracy of 90%, which is acceptable to the purpose of the work. However, it is decided to use either method to predict the menstrual date for users who newly registered or use the app for less than one-year period, as the utilized data set limitations. It is required to analyse more advanced seasonal level prediction models when the app is evolved with more users and collecting data.

Keywords: menstruation cycle, prediction, women

I. INTRODUCTION

A. Background

Menstruation is the shedding endometrium of the uterus that occurs once a month. This happens due to the hormonal effects on the body during puberty and menopause in women. Moreover, this is an individually unique cycle, and if this is systematic, it is considered one of the hallmarks of a woman's health. The process can be described as preparing the woman's body for pregnancy once a month, creating specialized tissue in the uterine wall to feed a fertilized egg, but removing previously created specialized tissue when the egg is not fertilized. Counting fertile times for pregnancy, using birth control methods to prevent unwanted pregnancies, and identifying pre symptoms about their body are the main procedures performed using the menstrual cycle. Also, due to the chemical reactions that take place in this process, the body undergoes physical and mental processes (Hillard, 2014).

Furthermore, based on the date of menstruation (after this known as date), it could understand if a woman's body is fertile. Therefore, women need to know the date of onset of menstruation to prepare their bodies physically and mentally and use the contraceptive methods required to conceive or prevent unwanted pregnancies. Hormone levels and body age, mood, diet, action and physical disorders, body weight, smoking, drug use, and stress are all factors that contribute to an irregular menstrual cycle, and the menstrual cycle can be used to diagnose preterm and gynaecological conditions. Hence, women need to know the approximate date (Bae, Park and Kwon, 2018).

B. Problem

The menstruation cycle is mainly based on the dimensionof dates, and those dates need to be counted and memorized. Nevertheless, it is not easy with day-to-day works of gender-specific works. Commonly women practice manual marking on the home calendar or mobile application. However, to predict the next date, many women and most mobile applications use 28 as a recurring cycle. Nevertheless, due to the individual cycle is dependent on the dynamic factors, the accuracy of the predictions becomes very low. Therefore, most women find it challenging to prepare physically and mentally and understand their reproductive health symptoms.

Therefore, the present work's base-work requires to development of a mobile application to provide the woman with a reproductive health plan based on the date. Then to predict the date, it required a better algorithm, and when searching the literature, it found that several works were carried out to develop an algorithm to predict the dateaccurately. However, most of them are based on evolutionary algorithms and machine learning. Apart from that, for predictions, there are various complex statistical models such as Seasonal Auto-Regressive Integrated Moving-Average (SARIMA) and Holt Winter's Exponential Smoothing (HWES).

Nevertheless, the present work is sufficient to have an 80% accurate prediction. Furthermore, through the initial discussion in base-work, it found that women need to know the exact date, but they, more importantly, look for cycle-related notices and guidelines, such as food and vitamins suggestions (Uthpala and Pradeep, 2020). Hence, rather than accurate prediction, it required individualistic cycle guidance from the future mobile application. For that, it required a straightforward methodology to suggest a prediction for of date.

C. Objective

This study is being done to find a satisfactorily accurate straight statistical model to find the pattern related to the menstrual cycle in women

II. LITERATURE REVIEW

Considering the algorithm to be used for the forecasting process, it can be considered that it is possible to proceed successfully by deriving a value given by a moving average. In predicting the date of menstruation, future predictions can be made by a moving average based on two dependent variables due to the presence of two simple data, the date of menstruation, and the length of menstruation. Parekh and Ghariya used three methods for this time series analysis named Simple Moving Average (SMA), Cumulative Moving Average (CMA), and Weighted Moving Average (WMA). And showed that the first two methods are best for data analysis in any application. In the case of a simple moving average, the prediction process is continued by finding the mean of the two segments and using that value to predict the value of the next segment. In their research, it used to find the mean between the menstrual days of the previous two months and use that average value to predict the menstrual day of the next month. When searching the cumulative moving average, it calculates the average value by making a cumulative of all existing data as it is based on the availability of the data (Parekh and Ghariya, 2015).

Hansun also evaluates the use of a moving average for future forecasting processes in periodic analysis. They offer five methods adding Exponential Moving Average (EMA), and Weighted Exponential Moving Average (WEMA) to the aforesaid three methods. Focusing on those methods, there is a drawback in the EMA, WEMA, and WMA, compared to SMA and CMA. That is, although all three methods, WMA, EMA, and WEMA, are derived from SMA, they add more weight to the new data values. Therefore, since the SMA and CMA methods have equal weight values, and do not affect other weight factors, the menstruation date prediction process can be done better, simpler, and more efficiently (Hansun, 2016).

Fattah evaluates the use of forecasting for identifying future demand by analysing the historical data. These data do have not any order. Because of that forecasting methods use to categories these data in the correct order. Automated Progressive Combined Movement Average is called ARIMA. Use for statistical and economic, and especially time series analysis and identify the behaviour of given data set. Model is a generalization ARIMA model. Automatic Progressive Movement Average (APMV) is the parallel methodology the same as ARIMA. Two of those modes are compatible with statistical information to perceive the information or predict the longer term of the series. ARIMA model is that the correct and best-suited methodology used for time prediction. The accuracy of the developed model was evaluated by scrutiny the experimental and the simulated amendment of the information within the same amount. Therefore, this model is often accustomed analyse and model the demand during this situation. The ARIMA procedure of the SPSS statistic module permits estimating the coefficients of the models that have antecedently known by providing the parameters p, q, and d, employing a quick most chance estimation algorithmic rule. ARIMA has been able to achieve successful results when compared to other methods. Because the release of high precision results has enabled the successful completion of several projects. Also, researchers recommend ARIMA for time series analysis and future forecasting (Fattah et al., 2018).

Karnaker, Halder, and Saeker introduce SMA as a method used to smooth out deviations, and variations over short periods, as well as to show conditions over long periods, and to use the forecasting process after obtaining the mean value after observing the entire dataset. And this responds and confirms the resistance, and support in every single vary. That is an advantage of SMA. Also, the CMA calculates the average value of all the data in a data stream (Karmaker, Halder and Sarker, 2017).

Numerous prognostication models are planned to search out an efficient methodology that will be applied to sensible things. These techniques principally deem complicated statistics, AI techniques, and enormous amounts of meteorological and topographical knowledge. Ideally, these ways minimize the chance of failure among the energy system and forecast its reliability by modelling or simulating future eventualities. The on the market prediction models may be classified into three main classes like qualitative techniques, quantitative techniques, and artificial neural networks. Qualitative techniques square measure supported professional opinion and private judgment. Quantitative techniques square measure supported mathematical models, which may be any classified as statistic or causative prognostication techniques. ARIMA is not a simple method. It requires more experience and knowledge to maintain the standard and provide quality outcomes. Causative prognostication is employed to spot relationships between dependent and freelance variables. the standard of causative prognostication models depends on the accuracy of the input factors. thanks to the high fluctuation of things poignant radiation, however, the provision and accuracy of those models are questionable. ARIMA is thought to be a modern technology, which applies once all the fairness information is lengthened and the correlation between past observations is stable. Many studies in the literature have used the ARMA and ARIMA models for radiation predictions. The ARMA and ARIMA models are additionally compared in terms of suitability values designed to enable log-probability to function. As a result, the best statistical models for prediction and the corresponding parameters can be determined in detail. Several possible comparisons are made for forecasting purposes. In the previous work, the forecast function of many models did not have an adequate and temporal sequence of information (Alsharif, Younes and Kim, 2019).

Apart from the utilized methods, the Auto-Regressive Integrated Moving Average (ARIMA) also has been identified as a powerful fundamental prediction model in statistical studies in the subject area.

The present work selected CMA and ARIMA which represent two prominent predictions used in the present discipline of the study as candidate methodologies to be automated in the app development.

III. METHODOLOGY

The present research proposed to predict the next date of a woman based on the average dates of the past data available. First, it has done a literature review to identify the suitable statistical models which can be used for time-series data prediction. Then to find the accuracy levels of different prediction models, it launched a field data collection. The resulted data set was analysed using two candidate models and make recommendations.

A. Data Collection

Then, a random sample of 30 women between the ages of 20 and 35 years was selected, and data were collected using a well-designed structured form. When this form was set up properly, several steps and versions followed the standard and structured the form. It was first sent to a peer batch mate to get feedback on the form's shortcomings and where it needed to be redone. Subsequently, the form was sent back to 3 other women to inspect the existing adjustments in the new version, redesigned in response to those responses, and to receive feedback. In this way, the correct new version of the form (as shown in Figure 1.0) was created based on the feedback received so that the form fillers could easily and clearly understand and sent to the selected sample of 30 women.

The sample of the women was instructed to fill out this form. Then in the month, the survey was started, the 30 women were contacted on a separate telephone conversation to find out the date of their menstrual period over the phone, and with the help of that date, the researcher updated a separate sheet (As shown in figure 2.0). The researcher also instructed the 30 women to fill the monthly report, including their menstrual dates. Throughout the year, both the forms were filled by each woman and the author. After doing so, the researcher performed an analysis of the accuracy of the dates in the updated forms. Subsequently, the classification process for the analysed data was carried out over the phone with the 30 women.

Monthly Data Entry Sheet				
Code Number				
Date (the first date of the cycle)				

Figure 1. Form filled by a woman (English Translation)

Monthly Data Entry Sheet				
Code Number				
Code Name	Tp. No			
DOB	Education			
Marital Status	Known Cycle length			
Date (the first date of the cycle)				

Figure 2. Form filled by the researcher (English Translation)

B. Data preparation

The first is to study the data set and removed such outliers. The removing considerations are failing to mark more than 20% of the dates, and women started to use birth control pills, drugs for menstrual cycle issues, and alcohol. Eighteen women in the sample failed to fill out the form for various practical reasons, such as suffering from illness during that month, personnel reasons, social needs, and ethical conditions such as losing close relationships where the author also loses control of acquiring data. Also, 10 out of the remaining 12 women successfully and systematically presented to the researcher a set of data that included the first day of their menstruation. After interviewing the other two again, it could obtain missing data. Furthermore, after collecting the datasets, the dates were sent back to the 12 women for certification and utilized in analysis with their consent.

Further, it found that abnormal cycles varying from 45 to 56. Then those outliers were normalized using equation 1.

$$CD_n = \left(\frac{CD_{n-1} - CD_{n+1}}{2}\right) + CD_{n-1} \quad \text{----- Equation 1}$$

Where CDn is the date of the abnormal cycle, n is the date.

C. Data Analysis

Initially, using the CMA (equation 2) and Auto Regression Integrated Moving Average (equation 3) the collected data were analyzed.

$$\mathrm{CMA}_{n+1} = \left(\frac{x_{n+1}+n\mathrm{CMA}_n}{n+1}\right) \quad \text{-----}$$
Equation 2

Where n in sequence number of data, x is data.

$$y_t = \varepsilon_t \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \quad \text{----- Equation}$$
3

To do the data analysis process using ARIMA model, initially, a time series plot was constructed to identify the nature of the menstruation cycle lengths of women and that is shown in Figure 3. And in there, menstruation cycle lengths of a particular women with WID 01 were considered. This is because from woman to woman, the menstruation cycle, which is caused by the hormonal effect on their body, shows a unique pattern to each other.

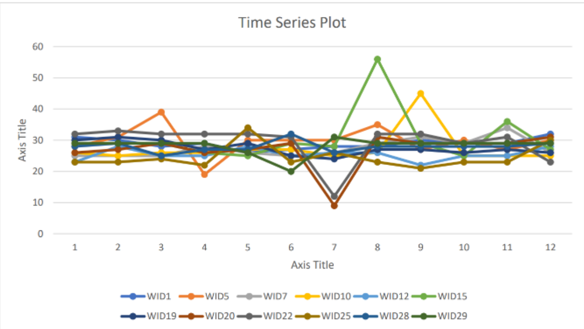

Figure 3. Time Series Plot time to identify the nature of the menstruation cycle lengths of women

And to do the analysis in that time series, used two functions named Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). They are initially similar to each other, but the two functions differ from each other in the process of incorporating or exluding the correlations in the calculations.

According to the ACF plot given in the following figure, up and down fluctuations can be identified. Also, autocorrelations of all lags are insignificant except the initial lag.

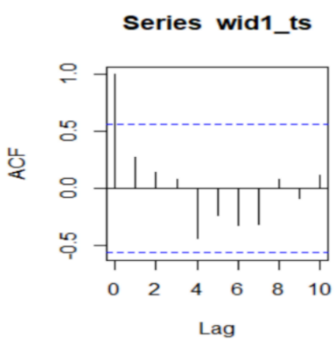

Figure 4. Autocorrelation Function plot

According to the PACF plot partial autocorrelations of all lags are insignificant.

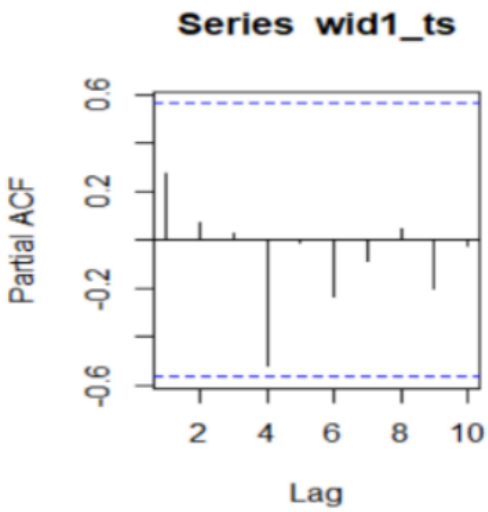

Figure 5. Partial Autocorrelation Function plot

In the checking process of the analysis, Stationarity of the original series was checked using Dickey-Fuller test.

Hypotheses for test the stationary of series:
H_0: Series is not stationary
H_1: Series is stationary

Table 1. Augmented Dickey-Fuller Test Data

Dickey-Fuller	Lag order	p-value
-3.3235	0	0.08841

According to the Table 1.0, p-value obtained by Augmented Dickey-Fuller Test is not less than 0.05 at 5% level of significance. It concludes that the original series is not stationary.

Since the original series is non-stationary the first difference series was considered.

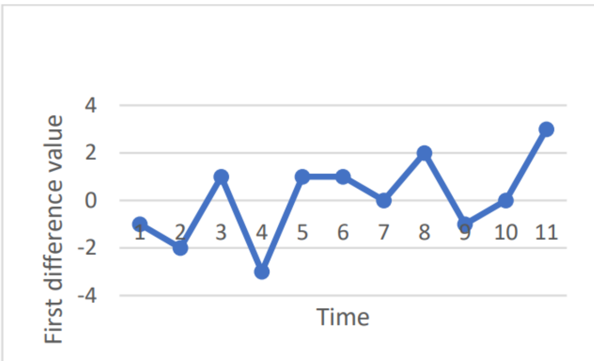

Figure 6. Non-Stationary Original Series

Stationarity of the first difference series was checked using Augmented Dickey-Fuller test.

Table 2. Augmented Dickey-Fuller Test Data

Dickey-Fuller	Lag order	p-value
-4.6522	0	0.01

According to the Table 2, p-value obtained by Augmented Dickey-Fuller Test is less than 0.05 at 5% level of significance. It concludes that the first difference series is stationary.

Since the first difference series is stationary and ACF plot depicts that the autocorrelation of all lags is insignificant except the first lag and further PACF plots depicts that the partial autocorrelation of all lags is insignificant, ARIMA (0,1,0) model was decided to develop.

So, summary of ARIMA (0,1,0) model can be defined as follows.

Table 3. Summary of ARIMA (0,1,0) model

sigma^2 estimated	log likelihood	aic
2.818	-21.31	44.61

Table 4. Box-Ljung Test Data

X-squared statistic	df	p-value
12.234	10	0.2697

According to above Table 4. p-value of the test statistics provided by Box-Ljung test is greater than 0.05, Ho: cannot be rejected at 5% level of significance. Therefore, it concludes that the fitted model is adequate.

And to analyse the appropriateness of a linear regression model use the Residual analysis. So, should do a Normality test for residuals.

Normality test for residuals:
Hypotheses for the Shapiro-Wilk normality test:
Ho: Data are normally distributed
H_1: Data are not normally distributed

Table 5. Anderson-Darling normality Test Data

AD statistic	p-value
0.21752	0.7936

According to the Anderson-Darling normality test result, P-value of AD statistics (= 0.7936) is greater than 0.05. Therefore, Ho is not rejected at 5% level of significance and concluded that the residuals are normally distributed.

Also, according to results of the residual analysis, it can be identified that the underlined distributional assumptions of residuals are satisfied by the fitted model. The verification the accuracy of the fitted mode is shown in Table 6.0 as a Summary of the ARIMA (0, 1, 0) model.

Similar procedure was carried out for menstruation cycle lengths of other women and the following results were shown in the Table 6.

Table 6. ARIMA prediction accuracies

Women ID	ME	RMSE	MAE	MPE	MAPE	MASE
Regular						
WID1	0.08591667	1.6073	1.252583	0.1168456	4.360667	0.9185611
WID7	0.41875	2.466452	1.91875	1.186714	6.50708	0.917663
WID19	-0.3939803	1.781854	1.374286	-1.682596	5.19745	0.889244
WID28	0.169	2.738625	1.835667	0.1094403	6.622261	0.9178333
Irregular						
WID5	0.08566667	7.664859	5.252333	-3.988727	20.34526	0.9170741
WID10	0.002166666	8.010413	3.8355	-2.869634	12.4637	0.9171848
WID12	0.1685833	2.549518	1.83525	0.194647	7.22495	0.917625
WID15	-0.08091667	12.04506	7.252417	-5.194676	20.93612	0.9169722
WID20	0.002166666	8.736898	4.668833	-13.32729	28.70533	0.9170923
WID22	-0.08066667	8.046744	3.919333	-8.578921	20.59257	0.9172908
WID25	0.001916666	4.966559	3.001917	-1.640782	11.6066	0.9172523
WID29	0.002416665	3.763873	1.83575	1.83575	7.001574	0.917875

According to the accuracy measurements given in the Table 6, accuracy measurements (RMSE, MAPE) are significantly small for the models fitted for the menstruation data of each woman with regular cycle length compared to the women with irregular cycle length.

IV. RESULTS AND DISCUSSION

A. Sample Size

The study has only 13 data instances for a single woman. As the studied works of literature proved the utilization of three data to accurate prediction. Further, for software development evaluations Pradeep and Wijesekera, (2012) mathematically proved that thirteen samples can be utilized to achieve more than 85% accuracy. Hence the present study satisfied with the sample size.

B. Sample data

Among the woman who participated in the study, 42% were married and 58% were unmarried. 50% of the sample have completed the advanced level education and 25% each completed either ordinary level education or territory education. The distributions of age and occupancy levels are shown in Figures 4, 5 and 6.

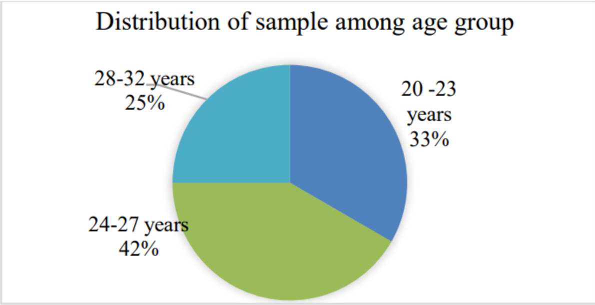

Figure 7. Age groups distribution of the sample

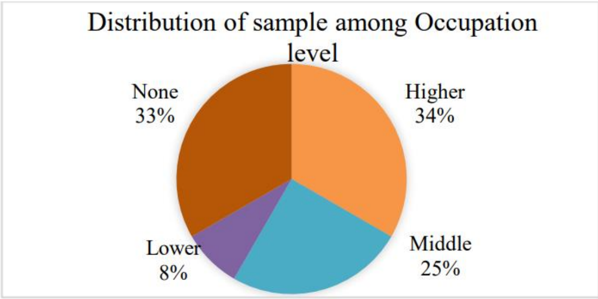

Figure 8. Occupation level distribution of the sample

As the base work is targeting to develop a mobile application for younger women, it can predict that the Higher 34% Middle 25% Lower 8% None 33% Distribution of sample among Occupation level 20 -23 years 33% 24-27 years 42% 28-32 years 25% Distribution of sample among age group selected sample is fit with the study and facilitate make suitable decisions.

C. Sample data

Table 7. CMA prediction accuracies

ID	Age	Occupation Level	Marital status	Average days of Mensural cycle	CMA Method	
					Average Variance of the prediction	Prediction Accuracy
1	22	Top	Yes	28	2	92.9%
2	22	No	No	28	2	92.9%
3	23	No	No	28	2	92.9%
4	23	Middle	No	28	1	96.4%
5	24	Middle	No	29	2	93.1%
6	24	Middle	No	30	2	93.3%
7	26	Top	No	26	3	88.5%
8	27	No	Yes	30	4	86.7%
9	27	Low	No	25	1	96.0%
10	28	Top	Yes	24	2	91.7%
11	32	No	Yes	25	1	96.0%
12	32	Top	Yes	28	2	92.9%
Average				27.4	2	92.8%

The CMA and results are shown in Table 7.0 and ARIMA results are shown in Table 5.0. Identically both the results show very high accuracy, i.e., 92.8% in CMA model and low error rate in ARIMA model consecutively. However, as shown in Figure 4.0, there is no collectively significant effect of utilization of amount of data to the prediction. When individually calculate the correlation between the number of cycles used and prediction variations, the coefficient average is 0.007, the median is -0.001 while the maximum value is 0.622 and the minimum is -0.584. Further, the result shows that the calculated average menses cycle (27.4 days) is in the commonly accepted range of 28.

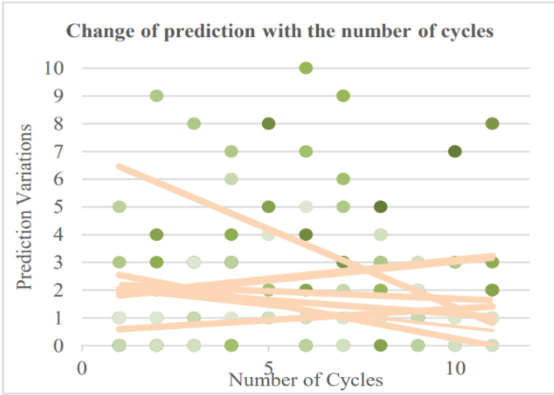

Figure 9. Number of cycles effect on the prediction accuracy

When compare the CMA and ARIMA results, the ARIMA result is much accurate than CMA. However, to calculate the ARIMA it needs to develop excessive software codes than the CMA and also ARIMA model is suitable for only regular menstruation patterns. As well, the prospective mobile app is planned to commence the calculate the mensural cycle using the value given by the user. Then it must provide prediction with the minimum usage of data. Hence, the app needs to change its cycle calculation method with time. Then it is suitable to utilize the simpler CMA method which utilizes fewer resources to develop and deploy at the initial stages.

V. DISCUSSION

Even though the menstrual cycle of a healthy woman is systematic, irregularities are inevitable. Hence prediction of the mensuration cycle needs to be realistic to the application, rather than the absolute accuracy. Then, the present work is required to select a substantial, simple, and straightforward method to calculate the next menstrual date for utilizing in the app development project.

A systematic field data collection was launched with 30 selected women but end up with twelve samples. However, the sample is appropriately distributed among the expected age group. Furthermore, rather than most of the availed data sets, present workable to administer to capture 13 events and 12 consecutive cycles. Therefore, the present research could evaluate the prediction accuracies for one year. Hence it considered the sample size is substantial for the research.

Among wide range of prediction models, then it used CMA and ARIMA methods which are popular fundamental timeseries statical methods. This is because the focus of this research is on identifying a menstruation cycle pattern that is unique from woman to woman, rather than identifying a menstrual cycle pattern that is common to all women. And those two methods are frequently used in the menstruation cycle calculation, the present work analysed the available data set. Both the methods provided resulted in high accuracies such as 92.8% (CMA) and low error rate in ARIMA model, showing confidence in the required mobile application development.

However, further analysis on the prediction accuracies found that the varying correlation between the number of cycles used to predict and the accuracy of the prediction from 0.6 to -0.6. Hence it can state that the software algorithm developers should pay more attention when utilizing either method of CMA or ARIMA to predict the individual women's menstrual cycle.

Starting from a simple method is more sustainable in software evolution as the management of the model is under the control of the programmer. Then the present work suggested utilizing the CMA, which is a less complex method to automate the menstrual cycle prediction, as the model needs to be customized in evolving passion.

REFERENCES

Bae, J., Park, S. and Kwon, J., 2018. Factors associated with menstrual cycle irregularity and menopause. pp.1–11.

Hansun, S., 2016. A Novel Research of New Moving Average Method in Time Series Analysis ISSN 2354-0082. (August 2014), pp.21–26.

Alsharif, M. H., Younes, M. K. and Kim, J. (2019) 'Time series ARIMA model for prediction of daily and monthly average global solar radiation: The case study of Seoul, South Korea', Symmetry, 11(2), pp. 1–17. doi: 10.3390/sym11020240.

Fattah, J. et al. (2018) 'Forecasting of demand using ARIMA model', International Journal of Engineering Business Management, 10, pp. 1–9. doi: 10.1177/1847979018808673.

Hillard, P.A., 2014. Menstrual suppression: Current perspectives. International Journal of Women's Health.

Karmaker, C.L., Halder, P.K. and Sarker, E., 2017. A Study of Time Series Model for Predicting Jute Yarn Demand: Case Study. 2017.

Parekh, P. and Ghariya, V., 2015. Analysis of Moving Average Methods. (1), pp.178–179.

Pirkalani, K.K. and Rad, Z.T., 2011. Menstruation cycle resembles parametric oscillation and subject to exact prediction: Mathematical model explaining hypothalamic behavior in response to stress, multiple pregnancies risk and external stimuli. European Psychiatry, [online] 26(S2), pp.1103–1103. Available at: <http://dx.doi.org/10.1016/S0924-9338(11)72808-5>.

Pradeep, R.M.M. and Wijesekera, N.T.S., 2012. Selecting a Usability Evaluation User Group – A Case Study the Development of a Hydro-GIS Tool Aiming Urban Flood Mitigation. In: Civil Engineering Research for Industry Symposium (CERIS) – 2012. [online] Moratuwa: Universiry of Moratuwa. Available at: <http://dl.lib.mrt.ac.lk/handle/123/9922>.

Uthpala, E. D. and Pradeep, R. M. M. (2020) 'Analysis on use of menstruation and fertility period tracker mobile applications', Student Symposium 2020-Faculty of Computing - KDU, p. 6. Available at: https://www.researchgate.net/publication/347746817_Analysis_on_use_of_menstruation_and_fertility_period_tracker_mobile_applications

ACKNOWLEDGMENT

The author would like to thank Dr. Lanka Dassanayake; specialist Obstetricians and

Gynaecologists, and Dr. EP Udeshika for imparting good knowledge on Obstetrics and Gynaecology fields. Further, the gratitude should go to Ms. HY Liyanage assistance in the review process of forms.

AUTHOR BIOGRAPHY

ED Uthpala is an undergraduate student in the Information Technology degree program at the General Sir John Kotelawala Defence University and will be graduating in 2021 with a BSc (Hons) in Information Technology. She is a member of IEEE and WIE. And she is interested in researching how information technology can be applied to the health sector.