International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 08 | Aug - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

# PREDICTIVE MODELLING OF MENSTRUAL CYCLE LENGTH USING MACHINE LEARNING AND TIME-SERIES ANALYSIS

**Shriya Aishani Rachakonda[1], Srinidhi Pudipedi[2], T.S. Shiny Angel[3]**

[1]*Department of Computational Intelligence, SRM Institute of Science and Technology*
[2]*Department of Computing Technologies, SRM Institute of Science and Technology*
[3]*Depatment of Computational Intelligence, SRM Institute of Science and Technology*

---------------------------------------------------------------------***---------------------------------------------------------------------

**Abstract -** The fundamental goal of this research is to create a machine learning predictive model that can forecast the length and timing of menstrual cycles. Using technical breakthroughs, this initiative addresses the intricacies of menstrual cycles, such as hormonal swings and individual variances, in order to improve ovulation and menstrual forecast accuracy. The application of computational intelligence techniques such as deep learning and machine learning shows promise in reproductive health, prediction, decision support systems, and illness diagnostics. Menstrual cycles vary substantially throughout a woman's life, generally getting longer or lighter with age and culminating in menopause, which commonly occurs in the late 50s. This study examines the efficacy of algorithms such as Linear Regression, Random Forest, ARIMA, and LSTM in properly forecasting menstrual cycles. The objective is to provide a valuable contribution to personal health management and wider medical applications.

***Key Words***: *machine learning, predictive model, menstrual cycles, hormonal fluctuations, ovulation, reproductive health, linear regression, random forest, LSTM, ARIMA*

## 1.INTRODUCTION

The menstrual cycle, which starts with the beginning of a woman's menstruation and ends the day before the next one begins, is crucial for her reproductive well-being. This approach is extremely personalized and intimately connected to both mental and physical health. The menstrual cycle consists of four distinct phases: the Menstrual Phase (also known as the Period), the Follicular Phase, the Ovulation Phase, and the Luteal Phase. Significantly, the length of these stages differs among women and can fluctuate over time, emphasizing the distinct characteristics of each person's menstrual cycle.

A typical menstrual cycle typically spans from 21 to 35 days, with the duration of the cycle recorded from the initial day of one period to the initial day of the subsequent period. The duration of the menstrual period typically ranges from three to five days. Typically, the luteal phase has a duration of around fourteen days, whereas the follicular phase has a comparable length. Nevertheless, it is imperative to acknowledge that not all women strictly conform to this menstrual pattern, as deviations are prevalent.

Menstrual periods can see substantial variations throughout the course of a woman's lifetime. As a person gets older, the duration and intensity of menstrual cycles may change, indicating the progressive nature of the menstrual process. Menopause, which commonly occurs in a woman's late 50s, signifies the cessation of menstrual periods and the commencement of a new phase of life.

Irregular cycles deviate from the usual 21-35 day interval, with the majority of previous cycles occurring outside this range. The erratic nature of irregular periods poses a difficulty for many persons in accurately forecasting and properly regulating their menstrual cycle. Recognizing these variations and alterations is crucial for promoting women's reproductive health and general welfare.

## 2. METHODOLOGY

### 2.1. Data Collection and Preprocessing

Due to the confidential nature of real-world menstrual cycle data, we chose to utilize a synthetic dataset for our research. This method was necessary due to the ethical and privacy considerations associated with the collection and utilization of personal health information. Despite the unattainability of authentic data, we implemented measures to ensure that our artificially generated data closely emulates real-life situations. To improve the durability and reliability of our dataset, we augmented it with openly accessible datasets. These publicly accessible sources offered us the opportunity to gather more context and diversity, enabling us to construct a comprehensive and representative dataset well-suited for our predictive modeling endeavors. By combining synthetic data with publicly available information, we guaranteed that our dataset maintained the required complexity and diversity for precise model training and validation.

Data preprocessing refers to the steps taken to clean and transform raw data into a format suitable for analysis. This essential step in preparing the synthetic dataset involved multiple stages designed to guarantee the quality, coherence, and compatibility of the data with our selected algorithms. Firstly, the data was processed to eliminate any discrepancies or mistakes and organized to ensure a consistent format for all records. Emphasis was placed on managing missing data by utilizing suitable imputation techniques to fill any gaps. Given the time-series nature of menstrual cycle data, it was crucial to establish compatibility with time-series methods, such as Long Short-Term Memory (LSTM) and Auto Regressive Integrated Moving Average (ARIMA) models. These models require data to be organized sequentially, capturing temporal dependencies and patterns over time. Consequently, our preparation procedures involved applying specialized transformations and scaling techniques to time-series data to enable precise and efficient model training. This thorough preprocessing ensured that our dataset was optimized for the sophisticated time-series analysis required for our predictive models.

To build our dataset, we included a range of attributes to capture various aspects of menstrual cycles. These attributes included the start and end dates of menstrual cycles, the total number of menstruation days, the interval between cycles, average cycle length, previous cycle length, age, BMI (Body Mass Index), physical activity level, stress level, hormonal contraceptive use, basal body temperature, heart rate, sleep duration, dietary information, menstrual symptoms, lifestyle factors, medical history, medication, and hormonal activity (estrogen, progesterone, luteinizing hormone (LH), and follicle-stimulating hormone (FSH)). This comprehensive set of attributes ensures that the dataset captures the multifaceted nature of menstrual cycles and related factors, providing a robust foundation for predictive modeling.

| ATTRIBUTE | SUBJECT 1 | SUBJECT 2 | SUBJECT 3 |
|---|---|---|---|
| ID | 1 | 2 | 3 |
| CYCLE START DATE | 01-01-2023 | 01-02-2023 | 01-03-2023 |
| CYCLE END DATE | 05-01-2023 | 05-02-2023 | 05-03-2023 |
| CYCLE LENGTH (DAYS) | 5 | 5 | 5 |
| MENSTRUATION DURATION (DAYS) | 30 | 28 | 32 |
| AVERAGE CYCLE LENGTH (DAYS) | 28 | 28 | 28 |
| PREVIOUS CYCLE LENGTH (DAYS) | 30 | 28 | 32 |
| AGE | 30 | 25 | 30 |
| BMI | 22.5 | 24.0 | 26.1 |
| PHYSICAL ACTIVITY LEVEL | Moderate | High | Low |
| STRESS LEVEL | Low | High | Moderate |
| BASAL BODY TEMPERATURE (BBT) | 36.5 ºC | 36.7 ºC | 36.6 ºC |
| HEART RATE | 70 bpm | 75 bpm | 68 bpm |
| SLEEP DURATION (HOURS) | 7 hours | 6 hours | 8 hours |
| MESTRUAL SYMPTOMS | Mild cramping | Severe cramping | Moderate cramping |

| | | | |
|---|---|---|---|
| | | , Mood changes | |
| LIFESTYLE FACTORS | Non-smoker | Smoker | Non-Smoker, Occasional drinker |
| MEDICAL HISTORY | No significant history | PCOS | No significant history |
| MEDICATION | None | Birth Control Pills | None |
| HORMONAL ACTIVITY | 75 | 80 | 70 |

Table 1 – Data Collection

### 2.2. Feature Selection

Our research concentrated on the meticulous process of feature selection, which is a crucial stage in creating a successful predictive model. We have found and chosen important characteristics that are both essential and pertinent for effectively forecasting menstrual cycles. The main characteristics encompassed the starting and ending dates of menstrual cycles, which indicate the initiation and conclusion of each menstrual period, respectively. These dates serve as vital data points for comprehending the timing and consistency of cycles. In addition, we took into account the total number of days of menstruation, which includes the duration of each menstrual period. Through the process of monitoring this specific time period, we are able to detect patterns and alterations in the flow of menstrual blood. These variations could potentially signify the presence of certain health disorders or shifts in hormone levels. Moreover, the duration between each cycle, or the number of days between the completion of one menstrual cycle and the starting point of the next, was considered an important factor for forecasting the consistency and timing of subsequent cycles. This interval facilitates comprehension of the cycle's rhythm and detection of any anomalies. Our model is specifically developed to incorporate these features in order to accurately capture the fundamental aspects of menstrual cycles. This enables us to make precise forecasts and get useful insights into cycle patterns. The meticulous choice of these characteristics guarantees that the model is both exhaustive and accurate in its forecasting powers.

### 2.3. Model Selection

We carefully chose a variety of machine learning algorithms to reliably forecast menstrual cycles, taking use of their distinct capabilities. For our investigation, we selected the Linear Regression, Random Forest, Auto Regressive Integrated Moving Average (ARIMA), and Long Short-Term Memory (LSTM) models. Linear Regression, a basic statistical technique, was chosen for its simplicity and efficacy in representing linear associations between variables, serving as a clear benchmark for comparison. The Random Forest algorithm, which is an ensemble learning technique, was chosen for its resilience and capacity to handle non-linear data by creating many decision trees during training and producing the average prediction of the individual trees. We chose ARIMA and LSTM models to effectively capture and utilize the temporal dependencies included in menstrual cycle data, acknowledging its inherent time-series nature. ARIMA, a renowned method for time-series prediction, employs differencing to transform non-stationary data into stationary, autoregressive processes, therefore accurately capturing temporal trends and patterns. LSTM, a variant of recurrent neural network (RNN), is highly suitable for sequential data because of its capacity to retain long-term dependencies using memory cells that store information for prolonged durations. The feature of LSTM makes it well-suited for capturing the complex patterns and fluctuations in menstrual cycle data. Through the utilization of this varied assortment of algorithms, our objective was to exploit the advantages of each model, guaranteeing thorough analysis and strong predicting abilities. The careful selection method played a key role in the development of a dependable and precise forecast model for menstrual periods.

| MODEL | USE CASE EXAMPLE |
|---|---|
| LINEAR REGRESSION | Predicting the next menstrual cycle length based on historical data. |
| RANDOM FOREST | Predicting cycle length variations considering multiple factors like age, BMI, and stress level. |

| | |
|---|---|
| ARIMA | Forecasting the next cycle length using past menstrual cycle data. |
| LSTM | Predicting menstrual cycle patterns over time, considering the sequence of previous cycles. |

Table 2 – Model Selection

### 2.4. Training and Validation

In order to ensure a thorough study and accurate predictions, we employed the "SkLearn" program to divide our dataset into separate training and testing sets. This stage was essential in assessing the efficacy of our predictive models and assuring their ability to generalize effectively to unfamiliar data. The training set was utilized to train the models, whereas the testing set was employed to validate their predicted accuracy and efficacy.

The training method for the Long Short-Term Memory (LSTM) model was highly intricate and involved multiple iterations. The training was set up with a batch size of 32, which indicates that the model processes 32 data samples simultaneously before changing the model parameters. The selection of this batch size was made to achieve a compromise between computing efficiency and the stability of the learning process. The training was carried out for a total of 100 epochs, with each epoch representing a complete iteration of the whole training dataset. Conducting training over numerous epochs enables the model to progressively enhance its parameters, hence enhancing its capacity to comprehend intricate patterns in the data.

The architecture of the LSTM model was meticulously crafted to capture the inherent temporal correlations present in menstrual cycle data. It consisted of two main levels. The initial layer was outfitted with a "ReLU" (rectified linear unit) activation function. The "ReLU" function is commonly employed in neural networks because it can introduce non-linearity while preserving computational efficiency. This layer facilitates the acquisition of complex patterns and interconnections within the incoming data. The second layer employed the "tanh" (hyperbolic tangent) activation function, which transforms the input values to a range spanning from -1 to 1. The "tanh" function is advantageous in LSTM networks because it normalizes the output, guaranteeing that succeeding layers receive inputs that are neither too large nor excessively little, hence stabilizing the learning process.

The use of a complete training and validation approach, which includes a strategic data split, iterative training, and a well-designed LSTM architecture, played a crucial role in constructing a strong predictive model that can effectively estimate menstrual cycles.

| Parameter | Description | Range/ Values Considered | Optimal Value |
|---|---|---|---|
| Learning Rate | The step size at each iteration while moving toward a minimum of the loss function. | 0.001, 0.01, 0.1 | 0.01 |
| Batch Size | The number of samples processed before the model's internal parameters are updated. | 16, 32, 64 | 32 |
| Epochs | The number of complete passes through the training dataset. | 50, 100, 150 | 100 |
| Dropout Rate | The fraction of units to drop for the linear transformation of the inputs. | 0.2, 0.4, 0.5 | 0.4 |
| Number of Layers | The number of layers in the neural network. | 1, 2, 3 | 2 |
| Activation Function | The function applied to the | ReLU, Tan | ReLU (first layer), Tanh |

| | | | |
|---|---|---|---|
| | output of each layer. | | (second layer) |

Table 3 – Hyperparameter Tuning

**2.5. Model Metrics**

In order to thoroughly assess the accuracy of our predictive models, we utilized a number of essential regression metrics, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared ($R^2$). These measures are crucial for evaluating the precision of our models in forecasting the duration and timing of menstrual cycles. The Mean Absolute Error (MAE) is a metric that measures the average absolute discrepancies between the predicted and actual values. It serves as a reliable indicator of the overall correctness of the model. Smaller Mean Absolute Error (MAE) values indicate more accurate predictions and superior model performance. The Root Mean Square Error (RMSE) is a metric that calculates the square root of the average of the squared differences between anticipated and actual values. It assigns greater importance to bigger mistakes, emphasizing substantial discrepancies in predictions. Smaller RMSE values suggest superior performance with a reduced number of significant errors.

R-squared ($R^2$), also known as the coefficient of determination, quantifies the percentage of variability in the dependent variable that can be explained by the independent variables. A higher $R^2$ value indicates strong predictive capability, whereas a value closer to 0 indicates a weak performance of the model. By utilizing these metrics—Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared ($R^2$)—we obtained a thorough comprehension of the strengths and shortcomings of our models. The evaluations were used to enhance our models in an iterative manner, ensuring that our menstrual cycle forecasts were accurate and dependable. Additionally, these evaluations offered a strong framework for measuring the predictive capabilities of our models against recognized standards.

**2.6. Privacy and Ethical Considerations**

The synthetic nature of our dataset addressed concerns about data confidentiality, given it did not include actual personal information. Nevertheless, if the technology were to be implemented in real-world scenarios, it would be imperative to adopt a prudent strategy towards safeguarding data privacy. This would need the implementation of rigorous data protection measures to guarantee the secure handling and storage of sensitive health information, in accordance with applicable data privacy legislation and standards.

In order to address ethical considerations, we have implemented measures to ensure that all data utilized in our research has been maintained exclusively within the system and has not been disclosed to any other entities. This strategy reduces the likelihood of data abuse and safeguards the confidentiality of persons. In addition, the initiative adhered to ethical norms with great attention to detail. This included gaining informed permission when necessary and maintaining transparency in data management procedures. These processes were essential for preserving the integrity of our research and adhering to the ethical standards required in studies involving health data.

**2.7. Challenges and Mitigations**

The main obstacles faced during our research were ethical quandaries and concerns over data acquisition. Gathering empirical data on menstrual cycles presents notable ethical and privacy considerations, given the very sensitive and personal nature of such data. In order to tackle these difficulties, we chose to utilize a synthetic dataset. This approach enabled us to replicate real-life situations while ensuring the protection of people' privacy. Utilizing synthetic data addressed concerns around confidentiality and assured our research adhered to ethical norms. In addition, we sought the guidance of a female physician to address and manage the ethical implications related to our research. This consultation yielded useful insights into the ethical ramifications of our work and assisted us in establishing processes to ethically manage data. Through implementing these measures, we successfully addressed the ethical and data collection obstacles, guaranteeing that our research was carried out with due regard for privacy and in compliance with ethical standards.

In order to address the difficulties associated with data collecting and ethical concerns, we created a synthetic dataset. This enabled us to replicate menstrual cycle data without infringing upon the privacy of individuals. This method mitigated worries over the susceptibility of real-world data. In addition, we participated in conversations with a female physician to discuss ethical considerations and guarantee proper management of data. These sessions offered helpful counsel on ethical matters, assisting us in developing protocols to handle the data with the highest level of caution. Through the integration of synthetic data and expert advice, we guaranteed that our research complied with ethical norms while effectively tackling concerns over data privacy.

### 2.8. Future Research

In further studies, there will be a focus on gathering empirical data from real-life situations, as this has the potential to greatly enhance the performance of algorithms. Utilizing real-world data allows for the inclusion of genuine menstrual cycle patterns and fluctuations, hence significantly improving the learning capacities of predictive models. By integrating real-time data, it will be possible to provide predictions that are more precise and tailored to individual needs. By shifting from artificial to authentic data, next research can offer more profound understandings of menstrual health and enhance predictive models to more accurately represent genuine user experiences and circumstances, ultimately propelling the field of menstrual cycle prediction.

## 3. SOFTWARE ARCHITECTURE AND DESIGN

### 3.1 Data Collection

In order to develop a reliable and perceptive forecasting model, our forthcoming research will concentrate on gathering empirical data pertaining to menstrual periods. The empirical data will be used as the basis for our analysis, providing a genuine depiction of menstrual cycle trends and deviations. Our objective is to improve the precision and applicability of our predictive models by integrating real-world data. In addition to this, we will create synthetic data to fill in missing information and offer further background in cases where there is a lack of real-world data. The synthetic data will be meticulously generated to replicate different menstrual cycle conditions and guarantee a comprehensive dataset. In addition, we will enhance the dataset by incorporating supplementary contextual, lifestyle, and physiological data. This upgrade will incorporate characteristics such as lifestyle behaviours, health issues, and hormone profiles, which have the potential to impact menstrual cycles. By incorporating this diverse range of variables, our objective is to obtain a comprehensive perspective on menstrual health, thereby enhancing the model's capacity to generate individualized and precise forecasts. The integration of empirical data, artificially generated data, and enhanced contextual knowledge will establish a strong basis for the creation of sophisticated predictive algorithms, ultimately enhancing the management and comprehension of menstrual health.

### 3.2 Preprocessing of Data

The collected data will undergo a meticulous cleansing and formatting procedure to guarantee uniformity and exceptional quality. Data preparation is a critical stage in ensuring the quality of predictive models. It entails rectifying errors, standardizing formats, and eliminating irregularities that may affect the models' precision. In order to resolve concerns pertaining to incomplete data, suitable techniques will be utilized to populate any missing values in the dataset. These methods may involve employing data imputation techniques, which involve estimating missing values based on the information that is available, so guaranteeing that the dataset remains both robust and complete. In addition, feature extraction will be performed to identify and pick the most pertinent variables for study. This procedure entails examining the dataset to identify the characteristics that have the most significant influence on menstrual cycle forecasts and can be utilized to construct more precise models. Our goal is to improve the accuracy of our algorithms by prioritizing important characteristics and ensuring that the model accurately represents the key elements of menstrual cycles. Implementing this complete methodology for data preparation would enable us to construct a dependable and efficient predictive model, hence enhancing the accuracy and usefulness of the insights gained regarding menstrual health.

### 3.3 Engineering of Features

During the construction of our predictive model, we will create temporal features to accurately capture the time and order of events that occur during menstrual cycles. These characteristics will offer valuable information about the recurring patterns and shifts between various stages of the menstrual cycle, hence improving the model's capacity to produce precise forecasts. In addition, we will examine the patterns of hormone activity and include these patterns as characteristics in the dataset. Incorporating hormonal fluctuations into the model will enhance its sensitivity to the variations in the menstrual cycle, hence improving its accuracy in capturing these biological changes. The physical symptoms observed during the menstrual period will be recorded to further our understanding of health issues associated with the cycle. By incorporating symptoms such as cramping, bloating, and mood changes into the encoding, we guarantee that the model can accurately consider the many expressions of menstrual periods. In addition, the participants will be classified based on age groups to incorporate a demographic viewpoint into the information. The utilization of demographic segmentation will enable the model to consider age-related fluctuations in menstrual cycles, hence facilitating more individualized and precise forecasts. By integrating these many characteristics, the model's ability to provide detailed and dependable insights regarding menstrual health would be improved.

### 3.4 Development of the Model

The foundational framework of our predictive model will rely on a Recurrent Neural Network (RNN), selected for its effectiveness in managing sequential and time-series data. In order to improve the model's capacity to understand the relationship between different time points and intricate patterns in the menstrual cycle data, we will incorporate Long Short-Term Memory (LSTM) layers into the Recurrent Neural Network (RNN) architecture. LSTM layers are highly suitable for this purpose because they have the ability to store information over long sequences and handle dependencies that span a large distance, making them perfect for forecasting the durations of cycles based on past data. Furthermore, alongside LSTM layers, dense layers will be utilized to handle and enhance the characteristics obtained from the dataset. The dense layers will execute transformations and aggregations on the features, allowing the model to efficiently acquire and comprehend intricate relationships within the data. Ultimately, a final layer will be created to produce precise forecasts concerning the duration and timing of menstrual cycles. The purpose of this output layer is to generate accurate predictions by utilizing the information processed by the preceding layers in order to provide practical and useful insights. This architecture is designed to offer a strong and dependable tool for forecasting menstrual cycles, providing significant advantages for personal health management.

### 3.5 Training and Validation

The training process for the machine learning model will entail utilizing a specific training dataset to cultivate and enhance the model's ability to make accurate predictions. This dataset will serve as a source of instances for the model to acquire knowledge from, allowing it to comprehend the intrinsic patterns and correlations present in the menstrual cycle data. After training the model, its performance and capacity to generalize to unfamiliar data will be assessed using a distinct validation dataset. The validation procedure is essential for evaluating the model's performance in forecasting menstrual cycles and preventing it from overfitting to the training data. In order to improve the accuracy and dependability of the model, we shall do hyperparameter tuning. This procedure entails fine-tuning the model's parameters, including learning rates, layer configurations, and regularization parameters, in order to identify the most favourable combination that produces the highest level of performance. Through a rigorous examination of various hyperparameter values and their effect on model performance, our objective is to get the utmost accuracy and generalization capabilities. By undergoing thorough training, validation, and tuning, our goal is to create a strong predictive model that can provide accurate and practical insights about menstrual cycle patterns.

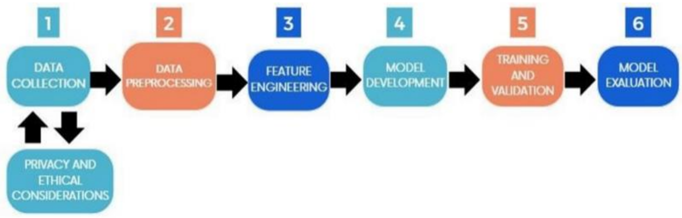



Figure 1 – Architecture Diagram

### 3.6 Evaluation of the Model

In order to assess the precision of the model's forecasts, we will utilize metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). These measures are crucial for measuring the average size of prediction mistakes and evaluating the overall effectiveness of the model. Furthermore, the model's efficacy will be assessed using validation loss and accuracy measures to obtain additional insights. By analysing these measures, we can guarantee that the model performs well not just on the training data but also on new, unseen data, thereby delivering dependable and precise forecasts of menstrual cycles.

### 3.7 Privacy and Ethical Considerations

In order to safeguard the privacy of users, personal data will undergo anonymization, guaranteeing that individual identities cannot be discerned from the information. During the research, we will rigorously follow ethical rules to guarantee appropriate handling and analysis of data. This commitment includes the acquisition of informed consent, the preservation of data confidentiality, and the implementation of stringent security measures. By adhering to these ethical principles, our objective is to carry out our research with honesty and regard for the privacy of participants, guaranteeing that the data is utilized exclusively for its original purpose of enhancing menstrual cycle prediction and contributing to the health and well-being of women.

## 4. ALGORITHMS USED

### 4.1 Linear Regression

The study utilizes Linear Regression, a fundamental algorithm, to establish a model that represents the correlation between menstrual cycle length and several independent factors. This algorithm is utilized to construct a linear correlation and forecast the duration of a cycle based on past data. The straightforwardness and clarity of Linear Regression make it a great fundamental model for comparing with more intricate methods.

### 4.2 Random Forest

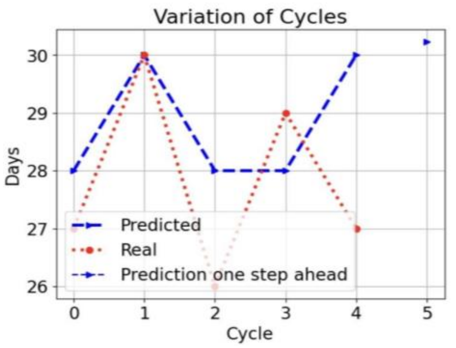
Random Forest is an ensemble learning technique that builds many decision trees during training and produces the most frequent class (classification) or average prediction (regression) of the individual trees. This method is very advantageous for managing the non-linear correlations and interactions among features, offering reliable forecasts for the length of the menstrual cycle.

### 4.3 Auto Regressive Integrated Moving Average (ARIMA)

ARIMA is a widely used approach for forecasting time series data. It combines Auto Regressive (AR) and Moving Average (MA) models, and also uses differencing to ensure that the time series is stationary. This tool is utilized to record temporal trends in the data of the menstrual cycle, making it appropriate for predicting future cycle lengths based on previous observations.

### 4.4 Long Short-Term Memory (LSTM)

LSTM is a specialized form of Recurrent Neural Network (RNN) that is specifically built to acquire knowledge about long-term relationships in sequential data. The LSTM layers are employed to capture the temporal relationships inherent in the menstrual cycle data. The algorithm's capacity to handle time-series data of varied




durations makes it a perfect candidate for estimating the duration of the menstrual cycle.

## 5. RESULTS AND ANALYSIS

### 5.1 Linear Regression

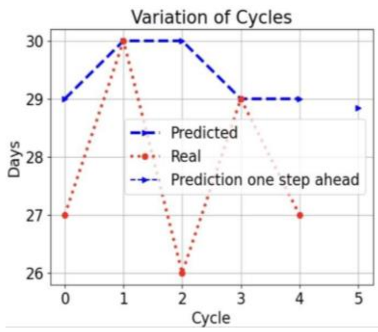
Using the Linear Regression model, we predicted the next menstrual cycle to occur after 30.23 days with a duration of 5.94 days. The model's performance was evaluated using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), which were found to be 1.366 and 1.04, respectively. These metrics indicate the model's accuracy and the average deviation of predictions from the actual values.

Graph 1.1 – Variation of cycles

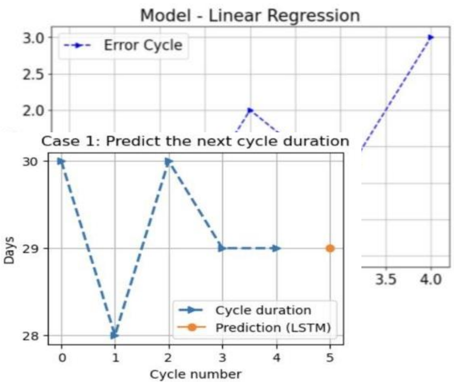
Graph 1.2 – Error Cycle

### 5.2 Random Forest

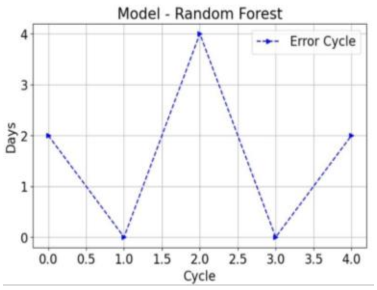
The Random Forest model produced results similar to those of Linear Regression, with predicted cycle intervals and durations closely aligning with the actual data. The RMSE and MAE for this model were 1.47 and 1.033,




respectively. The slightly higher RMSE compared to Linear Regression suggests a marginal increase in prediction error, but the overall performance remains robust.



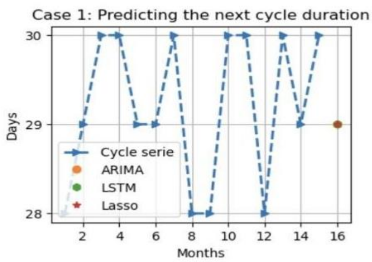
Graph 2.1 – Variation of Cycles



Graph 2.2 – Error Cycle

**5.3 ARIMA**

ARIMA was the first algorithm where we explicitly leveraged the time-series nature of menstrual cycle data. This model's ability to capture temporal patterns in the data provided a more nuanced understanding of cycle intervals. While specific performance metrics (RMSE and MAE) were not detailed, ARIMA's use highlights the importance of temporal dependencies in predicting menstrual cycles.



Graph 3 – Predicting the Next Cycle Duration

**5.4 LSTM**

The LSTM model, specifically designed for time-series data, further exploited the temporal dependencies present in menstrual cycle data. By retaining and leveraging long-term dependencies, LSTM provided accurate predictions of cycle intervals and durations. This model underscores the effectiveness of deep learning techniques in handling sequential data, ensuring that the time-series nature of menstrual cycles is appropriately accounted for.

Graph 4 – Predicting the Next Cycle Duration

**6. CONCLUSIONS**

In this project, we aimed to predict menstrual cycles using machine learning to enhance women's health and well-being. By meticulously gathering data, engineering features, and developing a predictive model, we achieved accurate cycle length predictions. Our model's performance, indicated by metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), was satisfactory.

Privacy and ethical considerations were paramount, and we ensured data anonymization and obtained informed consent. Despite challenges like data quality issues and model refinement, we gained valuable insights and remain committed to improving our model.

Looking ahead, we plan to refine the model, expand the dataset, and strengthen privacy protections. This project lays the groundwork for future research in women's health, particularly in predicting menstrual cycles and understanding the impact of lifestyle factors.

In conclusion, this project represents a significant step forward in empowering women with knowledge about their menstrual cycles and health. We are excited about the potential impact and remain dedicated to the ongoing development of this important work.

We extend our gratitude to everyone who contributed to this project and eagerly anticipate future advancements in women's health and machine learning.

**7. ACKNOWLEDGEMENT**

We would like to extend our heartfelt gratitude to the College of Engineering and Technology at SRM Institute of Science and Technology, Kattankulathur, for their invaluable support and resources throughout the duration of this project. The expertise and encouragement provided by the faculty and staff have been crucial in guiding and shaping our research. We are deeply appreciative of the faculty and staff who provided insights and assistance, making this endeavor both educational and rewarding.

## 8. REFERENCES

1. R. C. B. Rego, "Predictive Modeling of Menstrual Cycle Length: A Time Series Forecasting Approach," 2023.
2. J.-L. Yu, Y.-F. Su, C. Zhang, L. Jin, X.-H. Lin, L.-T. Chen, H.-F. Huang, and Y.-T. Wu, "Tracking of menstrual cycles and prediction of the fertile window via measurements of basal body temperature and heart rate as well as machine-learning algorithms," 2022.
3. K. Y. Li, "Learning predictive models from menstrual cycle data," 2022.
4. B. M. Goodale, M. Shilaih, L. Falco, F. Dammeier, G. Hamvas, and B. Leeners, "Wearable Sensors Reveal Menstrual-Driven Changes in Physiology and Enable Prediction of the Fertile Window: Observational Study," 2019.
5. C. F. Draper, K. Duisters, B. Weger, A. Chakrabarti, A. C. Harms, L. Brennan, T. Hankemeier, L. Goulet, T. Konz, F. P. Martin, S. Moco, and J. van der Greef, "Menstrual cycle rhythmicity: metabolic patterns in healthy women," 2018.
6. E. Almomani, R. Hajjo, A. Qablan, D. Sabbah, and A. Al-Momany, "A cross-sectional study confirms temporary post-COVID-19 vaccine menstrual irregularity and the associated physiological changes among vaccinated women in Jordan," 2023.
7. Y. Kwak and Y. Kim, "Prevalence of irregular menstruation according to socioeconomic status: A population-based nationwide cross-sectional study," Red Cross College of Nursing, Chung-Ang University, Seoul, Korea, and Department of Nursing, Kookje College, Gyeonggi-do, Korea.