**RESEARCH PROPOSAL**

[**Notes**: Acronyms used are detailed below the references section.]

**Research idea/topic:**  
Developing and Evaluating AI Models (ARIMA, LSTM, Prophet, XGBoost) for Predicting Menstrual Cycle Phases and Daily Symptoms in Individuals with Regular and Irregular Cycles.

**Study aim(s):**

1. To develop and compare the predictive accuracy of time-series AI models (ARIMA, LSTM, Prophet, XGBoost) for forecasting menstrual cycle start dates and symptom patterns.
2. To assess model performance differences between individuals with regular vs. irregular menstrual cycles.

**Research question:**  
Can AI models (specifically LSTM, ARIMA, Prophet, and XGBoost) accurately predict next-cycle start dates and daily symptom profiles for individuals with both regular and irregular menstrual cycles when trained on historical cycle length and self-reported symptom data?

**Study hypothesis:**  
***H₁:*** LSTM-based models will achieve significantly higher accuracy (RMSE < 1.5 days) in predicting cycle start dates compared to traditional statistical models (ARIMA), particularly for individuals with irregular cycles.  
***H₂:*** Models incorporating symptom data will outperform cycle-length-only models in predicting next-cycle start dates (AUC-ROC > 0.85).

**Literature Review**

Menstrual cycle irregularities affect 14-25% of reproductive-age women and are linked to conditions like PCOS and endometriosis. Current prediction apps (e.g., Clue, Flo) primarily use statistical averaging (28-day assumptions), with limited accuracy for irregular cycles (Epstein et al., 2020). While LSTM networks show promise in biomedical time-series forecasting (e.g., glucose prediction; Zhu et al., 2022), few studies apply advanced AI to menstrual data with symptom integration. Existing research lacks a direct comparison of ARIMA (suitable for stationary data) versus LSTM (handling long-term dependencies) in this context, especially concerning symptom-linked predictions (Chen et al., 2021). This gap is critical given the clinical relevance of cycle irregularity management.

**Methods**

**Study design:**  
*Observational cohort study* with prospective data collection (6 months) and retrospective model validation.

**Study population:**

* **Inclusion:**
  + 18-39 year-old menstruating individuals
  + Regular cycles (21-35 days ± 2-day variation) OR irregular cycles (<21 or >35 days ± ≥7-day variation)
  + Smartphone users (iOS/Android)
* **Exclusion:**
  + Pregnancy/lactation
  + Hysterectomy/oophorectomy
  + Hormonal IUD use
* *Target sample:* 600 participants (250 regular, 350 irregular)

**Sampling strategy:**

* *Stratified purposive sampling* via reproductive health clinics, *medical students*, and social media recruitment
* Oversampling irregular cycle cohort (PCOS, hypothalamic amenorrhea confirmed via medical records)

**Data collection and procedures:**

1. **Mobile App Platform:**
   * Custom app for daily logging:
     + Bleeding intensity (0-4 scale)
     + 20+ symptoms (e.g., cramping, mood, fatigue; Likert 1-5)
     + Lifestyle factors (sleep, stress)
2. **AI Modeling:**
   * **Features:** Historical cycle lengths, symptom severity vectors, age, and BMI
   * **Models:**
     + *ARIMA*: Baseline for stationary patterns
     + *LSTM*: 2-layer architecture with 64 units (capture long-term dependencies)
     + *Prophet*: Seasonality decomposition
     + *XGBoost*: Feature importance analysis
   * **Outcomes:**
     + Primary: Cycle start date prediction error (days)
     + Secondary: Symptom severity AUC-ROC (binary classification)
3. **Validation:**
   * 70/15/15 train/validation/test split
   * 5-fold cross-validation
   * Performance metrics: RMSE, MAE, AUC-ROC, F1-score

**Ethical considerations:**

1. **IRB Approval: [**Ideal] Protocol review
2. **Informed Consent:** Explicit opt-in for data sharing; withdrawal anytime
3. **Data Security:**
   * End-to-end encryption (HIPAA compliant)
   * Pseudonymization (user Ids, custom\_\_username vs. medical data)
   * AWS servers with ISO 27001 certification
4. **Privacy:**
   * No third-party data sharing
   * Symptom data aggregation for analysis
5. **Bias Mitigation:** Oversampling underrepresented groups (e.g., ethnic minorities, high BMI)

Key Implementation Notes:

1. **Irregular Cycle Definition:** Operationalized as >7-day variation in cycle length over prior 6 months (NIH criteria).
2. **Symptom Instrument:** Validated Menstrual Distress Questionnaire (MDQ) sub-items adapted for app.
3. **Model Tuning:** Bayesian hyperparameter optimization (e.g., LSTM learning rate: 0.001–0.1).

**References**

1. **Chen, T., Li, M., & Zhang, Y. (2021).**  
   *Comparative analysis of time-series models for predicting menstrual cycle patterns using mobile health data.* Computers in Biology and Medicine, 134, 104497.  
   DOI: 10.1016/j.compbiomed.2021.104497
2. **Epstein, D. A., Lee, N. B., Kang, J. H., et al. (2020).**  
   *Examining menstrual tracking to inform the design of personal informatics tools.* Proceedings of the ACM on Human-Computer Interaction, 4(CSCW), 1–22.  
   DOI: 10.1145/3415186
3. **Zhu, T., Li, K., Kuang, L., et al. (2022).**  
   *An LSTM-based model for personalized glucose prediction in type 1 diabetes management*. Journal of Biomedical Informatics, 128, 104308.  
   DOI: 10.1016/j.jbi.2022.104308
4. **WHO. (2023).**  
   *Menstrual health in human rights: A policy brief.* World Health Organization.  
   URL: https://www.who.int/publications/i/item/9789240080208
5. **National Institutes of Health (NIH). (2024).**  
   *Clinical guidelines for menstrual cycle variability.* NIH Office of Research on Women’s Health.  
   URL: https://orwh.od.nih.gov/research/clinical-research/clinical-guidelines
6. **Moos, R. H. (1968).**  
   *The development of a Menstrual Distress Questionnaire.* Psychosomatic Medicine, 30(6), 853–867.  
   DOI: 10.1097/00006842-196811000-00006

**Glossary of Acronyms and Technical Terms**

|  |  |  |
| --- | --- | --- |
| Acronym | Full Term | Explanation |
| AI | Artificial Intelligence | Computational systems performing human-like cognitive tasks |
| LSTM | Long Short-Term Memory | Recurrent neural network type for modeling long-range dependencies in sequential data |
| ARIMA | AutoRegressive Integrated Moving Average | Classical statistical model for time-series forecasting |
| XGBoost | eXtreme Gradient Boosting | Ensemble tree-based machine learning algorithm |
| RMSE | Root Mean Square Error | Measure of prediction errors (lower = better accuracy) |
| MAE | Mean Absolute Error | Average magnitude of prediction errors |
| AUC-ROC | Area Under ROC Curve | Measure of classification performance (0.5-1.0; higher = better) |

**Clinical/Medical Terms**

|  |  |  |
| --- | --- | --- |
| Acronym | Full Term | Explanation |
| PCOS | Polycystic Ovary Syndrome | A common endocrine disorder causing irregular cycles |
| IUD | Intrauterine Device | Contraceptive device (hormonal/non-hormonal) |
| EHR | Electronic Health Record | Digital patient health information system |
| MDQ | Menstrual Distress Questionnaire | Validated symptom assessment tool |
| NIH | National Institutes of Health | U.S. biomedical research agency |
| WHO | World Health Organization | UN global health authority |

**Methodological Terms**

|  |  |  |
| --- | --- | --- |
| Acronym | Full Term | Explanation |
| IRB | Institutional Review Board | Ethics committee overseeing human subjects research |
| HIPAA | Health Insurance Portability and Accountability Act | U.S. patient data privacy regulation |
| ISO | International Organization for Standardization | Body certifying data security standards |
| AWS | Amazon Web Services | Cloud computing platform |
| API | Application Programming Interface | Software communication protocol |
| BMI | Body Mass Index | Weight-to-height ratio (kg/m²) |

**Operational Definitions**

|  |  |
| --- | --- |
| Term | Meaning in Your Study |
| Regular cycles | Cycle length 21-35 days with ≤2-day variation |
| Irregular cycles | Cycle length <21 or >35 days with ≥7-day variation |
| Oversampling | Intentional over-representation of minority subgroups (e.g., high-BMI individuals) |
| Pseudonymization | Data de-identification technique replacing identifiers with artificial codes |

1. **Model-Specific Terms**
   * **Prophet**: *Not an acronym* - Facebook's time-series forecasting model
   * **Bayesian optimization**: Probabilistic hyperparameter tuning method
2. **Contextual References**
   * **GDPR** (General Data Protection Regulation): EU data privacy law (cited in ethics refs)