**INTRODUCTION**

Maternal health during pregnancy is pivotal not only for the well-being of mothers but also for the healthy development of the fetus and successful outcomes of pregnancy. Globally, complications during pregnancy and childbirth are leading causes of mortality and morbidity among women of reproductive age. The World Health Organization estimates that approximately 295,000 women died during and following pregnancy and childbirth in 2017 alone, with a significant majority of these deaths occurring in low-resource settings and being largely preventable​[1]. Pregnancy is a complex physiological process that spans approximately 40 weeks from the Last Menstrual Period (LMP) and is typically divided into three trimesters, each characterized by specific developmental milestones. The first trimester (0–13 weeks) is crucial for organ formation, the second trimester (14–26 weeks) focuses on growth and maturation, and the third trimester (27–40 weeks) prepares the fetus for birth. Despite these structured stages, pregnancy can bring about various complications such as high blood pressure, gestational diabetes, preeclampsia, preterm labor, and infections, which may adversely affect maternal and fetal health[2].

These complications underscore the unpredictability and challenges associated with maternal health. Even women in good health before pregnancy are at risk of developing complications that may escalate to life-threatening conditions if not promptly addressed. This reality has prompted an increased focus on predictive healthcare, leveraging data-driven approaches to anticipate and mitigate risks associated with pregnancy..

With the integration of machine learning (ML) into healthcare, new avenues have opened up for enhancing maternal health outcomes. ML models are increasingly being applied to predict pregnancy complications, offering the potential to revolutionize prenatal care by enabling early detection and intervention. These models utilize various types of data, including clinical parameters, patient histories, and real-time health data, to identify patterns that may predict complications such as pre-eclampsia, gestational diabetes, and intrauterine growth restriction​[1][3].

**PROBLEM STATEMENT**

Despite the promise shown by machine learning in improving predictions of pregnancy-related complications, several challenges impede their practical utility. First, the generalizability of these models across different populations remains limited due to the diversity in demographic and health-related characteristics[1]. Second, many studies suffer from small sample sizes, particularly in conditions like systemic lupus erythematosus (SLE), where the pool of pregnant patients is small. This often results in models that perform well in controlled settings but fail to generalize to the broader population[4]​.

Furthermore, the quality of data and the selection of features significantly impact the accuracy of predictions. Inaccurate or biased data can lead to erroneous predictions, which can have serious implications for patient care. Additionally, many models do not perform uniformly across all complications, showing variability in predictive accuracy depending on the specific condition being predicted[5]. This research proposes the development and implementation of an XGBoost Classifiermodel to predict pregnancy-related complications. **XGBoost Classifier** networks, are specifically adept at capturing sequential dependencies and temporal patterns in data, which are critical in modeling pregnancy-related health trajectories over time. Unlike traditional machine learning models that rely on static features they can process sequential data, such as time-series records of vital signs, biomarkers, and clinical observations.[1]

**AIM AND OBJECTIVES**

The aim of this thesis is to develop an XGBoost Classifier model to predict a wide range of pregnancy complications with enhanced accuracy, generalizability, and clinical utility.

Specific objectives include**:**

1. To investigate the interpretability of XGBoost Classifier models with attention mechanisms in identifying critical temporal features affecting pregnancy complications.
2. To develop an XGBoost Classifier model that can predict pregnancy complications in expectant mothers.
3. To validate the model across diverse populations to ensure its applicability in different healthcare settings.

**SIGNIFICANCE OF THE STUDY­­­­**

The significance of developing a data-driven machine learning model to accurately predict pregnancy complications cannot be overstated.

In many parts of the world, maternal and neonatal health outcomes remain suboptimal, with preventable complications leading to high rates of morbidity and mortality. Leveraging advanced modeling techniques, healthcare providers and public health authorities could identify at-risk pregnancies at a much earlier stage than is currently feasible. This early detection would not only facilitate timely and targeted clinical interventions, such as more frequent prenatal monitoring, personalized treatment regimens, and the application of preventive measures, but also enable more efficient use of limited healthcare resources, ultimately improving patient outcomes.

Moreover, an effective model would address critical limitations that characterize current prediction tools, which are often constrained by small sample sizes, imbalanced data, or a lack of interpretability. The enhanced specificity and sensitivity derived from a rigorously trained and validated machine learning system would thus contribute to safer pregnancies by minimizing delays in diagnosing and treating life-threatening conditions. In addition, the global application of this methodological framework stands to benefit maternal and neonatal healthcare across diverse populations and resource settings.

As such, the widespread adoption of these techniques could bolster healthcare infrastructures, inform policy decisions, and catalyze international collaborations aimed at improving maternal-child health. Over time, these initiatives would help reduce health disparities, promote equity in healthcare access, and significantly elevate the quality of life for mothers and their children worldwide.

**DEFINITION OF TERMS:**

Here are concise definitions of key terms used throughout this research:

1. **XGBoost Classifier (eXtreme Gradient Boosting):** An advanced gradient-boosting algorithm that uses decision trees to provide high performance and efficiency for both classification and regression tasks.
2. **SVM (Support Vector Machine)**: A supervised learning model used for classification and regression tasks that constructs a hyperplane or set of hyperplanes in a high-dimensional space.
3. **BiLSTM (Bidirectional Long Short-Term Memory):** A type of recurrent neural network that processes data in both forward and backward directions to better capture time-dependent information.
4. **Machine Learning**: The science of getting computers to learn and act like humans do, improving their learning over time in autonomous fashion by feeding them data and information in the form of observations and real-world interactions.
5. **Deep Learning**: A subset of machine learning involving algorithms inspired by the structure and function of the brain called artificial neural networks.
6. **Predictive Modeling**: The use of statistical techniques to predict future outcomes based on historical data.
7. **AUC-ROC (Area Under the Curve - Receiver Operating Characteristics)**: A statistical measure used to evaluate the performance of a classification model at various threshold settings.
8. **Precision**: The ratio of correctly predicted positive observations to the total predicted positives.
9. **Recall (Sensitivity)**: The ratio of correctly predicted positive observations to all actual positives.
10. **F1-Score**: A weighted average of precision and recall, providing a single score that balances both the concerns of precision and recall in one number.
11. **Cross-Validation**: A technique for assessing how the results of a statistical analysis will generalize to an independent data set, primarily used in settings where the goal is prediction.
12. **SMOTE (Synthetic Minority Over-sampling Technique)**: An algorithm that increases the number of cases in your dataset in a balanced way by making synthetic copies of the minority class

**LITERATURE REVIEW**

**1. "Using Machine Learning to Predict Complications in Pregnancy: A Systematic Review" by Bertini et al.[1]**

Bertini et al. conducted a systematic review to analyze the applicability and performance of machine learning (ML) techniques for predicting pregnancy complications. This review synthesized findings from 31 studies selected from an initial pool of 98 articles, focusing on studies published between 2015 and 2020. The reviewed studies predominantly utilized electronic medical records (48%), medical images (29%), and biological markers (19%) as input data, reflecting the diversity of data sources in obstetric research. The review highlighted that most ML models focused on predicting complications such as pre-eclampsia and prematurity, with the highest precision achieved using support vector machines (SVMs) and XGBoost for specific complications, reporting accuracies of 95.7% and 99.7%, respectively.

The systematic review used the PRISMA method to ensure rigorous selection and evaluation of the articles. Performance metrics such as AUC-ROC, sensitivity, and specificity were commonly used to validate the models, showcasing the high standards maintained in the field. However, the review identified several challenges, including the interpretability of black-box models, data imbalance, and limited generalizability due to single-center datasets. Furthermore, many models failed to address the scalability required for real-world clinical integration, limiting their practical applications. The authors emphasized the importance of developing explainable and user-friendly ML models, particularly for healthcare providers, who often lack the technical expertise to interpret complex algorithms. This review serves as a foundational study, advocating for more collaborative efforts to create scalable, interpretable, and clinically validated ML solutions for pregnancy care​.

**2. "Deep Learning-Based Feature Engineering to Analyze Maternal Health During Pregnancy" by Raza et al.[5]**

Raza et al. proposed a novel machine learning framework to predict maternal health risks during pregnancy by leveraging deep learning and feature engineering. Their approach utilized the DT-BiLTCN model, an innovative architecture combining decision trees, bidirectional long short-term memory (BiLSTM), and temporal convolutional networks (TCN). The model was tested on a dataset of 1,218 samples collected through IoT-based maternal health monitoring systems. The researchers resolved data imbalance issues using the synthetic minority oversampling technique (SMOTE) and enhanced the model's performance through rigorous hyperparameter tuning.

The DT-BiLTCN model demonstrated remarkable predictive capabilities, achieving an accuracy of 98% when paired with support vector machines (SVMs) for classification. Key predictors identified included diastolic and systolic blood pressure, heart rate, and maternal age, which were derived through exploratory data analysis. By integrating BiLSTM and TCN for temporal data modeling, the proposed architecture captured complex, non-linear relationships in time-series data, significantly improving prediction accuracy compared to traditional models.

Despite its strengths, the study acknowledges limitations. The reliance on IoT-generated data raises questions about scalability in regions lacking advanced technological infrastructure. Additionally, the model's complexity could hinder its adoption in clinical settings, where simpler models might be preferred. The study’s findings emphasize the transformative potential of deep learning in maternal healthcare while advocating for further research to enhance the model's accessibility and clinical utility​.

**3. "Machine Learning Models for Predicting Adverse Pregnancy Outcomes in Pregnant Women with Systemic Lupus Erythematosus" by Hao et al.[4]**

Hao et al. investigated the use of machine learning to predict adverse pregnancy outcomes in women with systemic lupus erythematosus (SLE), a condition associated with increased maternal and fetal risks. The study employed a small dataset comprising 51 women with 288 clinical variables. Six machine learning models, including random forest, SVM, and multi-layer perceptron, were evaluated. The random forest model emerged as the best performer, achieving high predictive accuracy even with limited sample sizes.

One notable aspect of the study was its use of real-time predictive models tailored to different gestational periods. The research demonstrated the importance of capturing temporal changes in pregnancy by evaluating the models’ predictive performance at various stages, including pre-pregnancy, first trimester, and third trimester. This chronological approach enabled the identification of critical risk factors that vary over time, offering a dynamic perspective on pregnancy risk assessment.

However, the study faced significant limitations. The small sample size and focus on a single ethnic group (Chinese Han) restricted the generalizability of the findings. Moreover, the study did not standardize feature selection processes, complicating the replication of results in different settings. Despite these challenges, the research highlights the potential of ML in addressing rare but high-risk conditions, paving the way for more personalized and adaptive approaches to pregnancy care​.

**4. "AI-Driven Predictions of Preeclampsia: A Meta-Analysis" by Gómez-Jemes et al.[3]**

This meta-analysis reviewed the performance of artificial intelligence (AI) models in predicting preeclampsia, a leading cause of maternal and neonatal morbidity. The analysis synthesized findings from multiple studies, comparing the effectiveness of AI models, including SVMs, random forests, and deep learning methods. The authors found that SVMs consistently outperformed other models in terms of AUC-ROC, with values exceeding 0.90 in most cases. The meta-analysis highlighted the use of diverse data types, including clinical parameters, biomarkers, and imaging data, to enhance prediction accuracy.

One key finding was the importance of integrating multiple data sources to capture the multifactorial nature of preeclampsia. However, the authors noted significant challenges, such as data imbalance and the limited interpretability of deep learning models. The review also emphasized the need for standardized metrics and evaluation protocols to facilitate comparisons across studies. By advocating for explainable AI and collaborative efforts, the authors underscored the importance of bridging the gap between research and clinical application, ensuring that AI-driven tools meet the needs of healthcare providers​.

**RELATED WORKS**

**1. "Machine Learning to Predict Pre-Eclampsia and Intrauterine Growth Restriction in Pregnant Women" by Gómez-Jemes et al.**

Gómez-Jemes and colleagues developed a machine learning model to predict pre-eclampsia and intrauterine growth restriction (IUGR), focusing on biomarkers and Doppler measurements of the uterine artery. The study employed multi-label classification to capture the simultaneous occurrence of these complications, demonstrating a nuanced approach to obstetric risk prediction. Among the classifiers tested, decision trees outperformed others, achieving high recall and AUC-ROC scores. The study’s primary limitation was its small sample size, which may affect the reliability of the results. Additionally, the dataset’s single-center origin restricts the model’s applicability across diverse populations. Despite these limitations, the research highlights the potential of ML to identify high-risk pregnancies and facilitate early intervention strategies​.

**2. "Predictive Performance of Machine Learning-Based Methods for the Prediction of Preeclampsia" by Melinte-Popescu et al.**

Melinte-Popescu et al. conducted a prospective study evaluating the predictive performance of machine learning models in identifying preeclampsia. The study used first-trimester clinical data and tested models such as decision trees, SVM, naive Bayes, and random forests. Decision trees and SVM showed exceptional performance for early-onset preeclampsia, while naive Bayes and random forests performed well across all types. Limitations include challenges in handling moderate and severe cases of preeclampsia and scalability to larger datasets. Nevertheless, the study underscores the potential of machine learning in enhancing early detection and management of preeclampsia, providing valuable insights for clinical adoption​.

The table below summarizes key points from each study, framing them within the context of their contributions to the field and noting their limitations.

**SUMMARY OF REVIEWED WORKS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Title of Paper/Year** | **Problem Addressed** | **Methodology** | **Summary of Result** | **Limitation** |
| **Predicting Adverse Pregnancy Outcomes Using ML Models for SLE Patients** (Hao et al., 2023) | Prediction of adverse outcomes in SLE pregnancies | Used a dataset of 51 pregnant women and applied six ML models. | RF showed best performance with real-time predictive capabilities across different gestational stages. | Small sample size limits generalization of results. |
| **Using Machine Learning to Predict Complications in Pregnancy: A Systematic Review** (Bertini et al., 2022) | To summarize ML techniques for predicting perinatal complications | Systematic review of 31 studies from databases like PubMed. | SVM showed high precision in predicting prematurity and neonatal mortality. | Generalizability issues due to varying study designs and populations. |
| **Ensemble Learning-Based Feature Engineering for Maternal Health Risk** (Raza et al., 2022) | Enhancing prediction of maternal health risks | Applied ensemble learning and deep neural networks to maternal health data. | Achieved high accuracy with SVM post-feature engineering using DT-BiLTCN. | Complexity of models may hinder real-world applicability. |
| **Machine Learning for Predicting Preeclampsia and Growth Restriction** (Gómez-Jemes et al., 2022) | Early detection of pre-eclampsia and intrauterine growth restriction | Utilized Doppler measures and ML classifiers on a dataset of 95 women. | Decision tree variant excelled in predicting with an AUC of 0.87. | Model struggled with distinguishing between types of conditions without clinical distinctions. |
| **Robust ML Model for Maternal Health Risk Prediction** (Pawar et al., 2022) | Prediction of maternal health risks. | Traditional ML algorithms tested on health datasets. | Proposed a more robust model that handled variable scenarios better with 70.21% accuracy. | Performance might still be suboptimal in less-controlled, real-world environments. |

CHAPTER 3

**METHODOLOGY**

**Introduction**  
This chapter provides a comprehensive, step-by-step methodology for developing a sequence-oriented machine learning model specifically, an XGBoost Calssifier network, to predict pregnancy complications in expectant mothers. The overarching aim is to leverage the temporal dynamics of maternal and fetal health indicators over the gestational period, thereby enabling the early identification of risks and facilitating timely clinical interventions.

**Introduction**

Pregnancy is a transformative journey with significant physiological and emotional changes for expectant mothers. These changes, while natural, can lead to complications that pose risks to both the mother and fetus. Predicting these risks early is pivotal to ensuring timely intervention and better outcomes. This methodology outlines an elaborate and systematic approach for predicting pregnancy complications using machine learning and statistical modeling, ensuring replicability and clinical relevance.

**1. Dataset Acquisition and Understanding**

**1.1 Data Sources**

The dataset comprises clinical and demographic information collected from hospitals, clinics, and maternal health monitoring systems. Features include:

* **Age:** Maternal age during pregnancy.
* **SystolicBP/DiastolicBP:** Blood pressure values in mmHg.
* **BS (Blood Sugar):** Glucose levels measured in mmol/L.
* **BodyTemp:** Body temperature in Fahrenheit.
* **HeartRate:** Resting heart rate in beats per minute.
* **RiskLevel:** Target variable categorizing pregnancy risks as low, mid, or high.

**1.2 Initial Data Exploration**

1. **Data Overview:**
   * Load and examine the dataset using **.info()**and .**describe()** to identify feature types, missing values, and summary statistics.
   * Visualize initial patterns through count plots, histograms, and box plots.
2. **Duplicate Handling:**
   * Identify duplicate rows using .duplicated().
   * Drop duplicates to ensure data quality and integrity.
3. **Outlier Detection:**
   * Use box plots and z-scores to detect anomalies in features like HeartRate and BS.
   * Validate whether detected outliers are biologically plausible or data entry errors.
4. **Feature Distribution:**
   * Evaluate distributions of features like BS and Age to detect skewness.
   * Apply transformations (e.g., log or Box-Cox) to normalize skewed data for better model performance.

**2. Data Preprocessing**

**2.1 Cleaning and Preparation**

1. **Imputation:**
   * Address missing data using mean/median imputation or predictive modeling for continuous variables.
2. **Encoding:**
   * Encode categorical variables, such as RiskLevel, using label encoding or one-hot encoding.
3. **Feature Scaling:**
   * Apply StandardScaler to normalize continuous features, ensuring consistent scales across variables.
4. **Feature Engineering:**
   * Derive new features (e.g., BMI categories) if additional variables are available.

**2.2 Train-Test Split**

* Split the dataset into training (70%) and testing (30%) subsets using stratified sampling to preserve the distribution of RiskLevel.

**3. Exploratory Data Analysis (EDA)**

**3.1 Univariate Analysis**

* Assess the distribution of individual features using histograms, density plots, and count plots.

**3.2 Bivariate Analysis**

* Explore relationships between features and the target variable (RiskLevel):
  + Box plots for numerical features (e.g., BS, Age) versus RiskLevel.
  + Scatter plots for correlations (e.g., Age vs. BP).

**3.3 Correlation Analysis**

* Generate a correlation matrix to identify multicollinearity among predictors.
* Use techniques like Variance Inflation Factor (VIF) to remove redundant features.

**4. Statistical Hypothesis Testing**

**4.1 Skewness Analysis and Transformation**

* Identify skewed features using .skew() and apply transformations (e.g., log, Box-Cox).

**4.2 Hypothesis Testing**

* Define null and alternative hypotheses:
  + **H0:** RiskLevel does not vary by Age.
  + **H1:** Women aged 30+ have higher RiskLevel compared to younger women.
* Use chi-square tests for categorical comparisons and t-tests for mean differences.

**5. Model Development**

**5.1 Machine Learning Models**

1. **Baseline Models:**
   * Logistic Regression for initial benchmarking.
2. **Advanced Models:**
   * K-Nearest Neighbors (KNN): Non-linear relationships.
   * Support Vector Machine (SVM): High-dimensional boundaries.
   * Random Forest (RF): Feature importance and robustness.
   * Gradient Boosting (GBM) and XGBoost: Iterative optimization for better accuracy.

**5.2 Model Tuning**

* Perform hyperparameter tuning using GridSearchCV for each model:
  + Random Forest: Number of estimators, max depth.
  + XGBoost: Learning rate, max depth, min child weight.

**6. Evaluation Metrics**

**6.1 Classification Metrics**

* Accuracy, Precision, Recall, F1-Score.
* Confusion matrix to assess performance across risk categories.
* ROC-AUC for model discrimination.

**6.2 Statistical Metrics**

* p-values and confidence intervals for hypothesis validation.
* Adjusted R-squared for regression models.

**Columns in the dataset:**

* Age: Age in years when a woman is pregnant.
* SystolicBP: Upper value of Blood Pressure in mmHg, another significant attribute during pregnancy.
* DiastolicBP: Lower value of Blood Pressure in mmHg, another significant attribute during pregnancy.
* BS: Blood glucose levels is in terms of a molar concentration, mmol/L.
* HeartRate: A normal resting heart rate in beats per minute.
* Risk Level: Predicted Risk Intensity Level during pregnancy considering the previous attribute.

**CONCLUSION**

This research targets the critical challenge of predicting pregnancy-related complications through a mixed-methods approach, combining advanced machine learning techniques with insights from healthcare professionals. The methodology revolves around deploying an XGBoost Classifier model, known for its efficacy in analysing complex, non-linear patterns prevalent in medical time-series data.

The essence of this study is not merely in the technical execution but in bridging the gap between data-driven predictions and clinical applications. Through rigorous data pre-processing, model development, and validation processes, we aim to establish a robust predictive model. This model is scrutinized using comprehensive performance metrics like accuracy, precision, recall, F1-score, and AUC-ROC to ensure their effectiveness and reliability in clinical settings. Significantly, this research emphasizes the integration of qualitative insights, ensuring that the predictive models align well with the practical realities and nuanced needs of maternal healthcare. This approach fosters a cycle of continuous improvement and adaptation, reflecting the dynamic nature of healthcare delivery. By engaging with healthcare providers for feedback and continuously updating the models with new data and insights, we ensure that our solutions remain relevant and effective over time.

Ultimately, by enabling earlier detection and more precise management of pregnancy-related complications, this research seeks to contribute to safer pregnancies and healthier outcomes for mothers and children. It underscores the transformative potential of integrating cutting-edge technology with clinical expertise to enhance healthcare delivery and patient outcomes.

**REFERENCES**

[1] A. Bertini, R. Salas, S. Chabert, L. Sobrevia, and F. Pardo, “Using Machine Learning to Predict Complications in Pregnancy: A Systematic Review,” *Front. Bioeng. Biotechnol.*, vol. 9, p. 780389, Jan. 2022, doi: 10.3389/fbioe.2021.780389.

[2] A. Baylis, W. Zhou, E. Menkhorst, and E. Dimitriadis, “Prediction and prevention of late-onset pre-eclampsia: a systematic review,” *Front. Med.*, vol. 11, p. 1459289, Nov. 2024, doi: 10.3389/fmed.2024.1459289.

[3] L. Gómez-Jemes, A. M. Oprescu, Á. Chimenea-Toscano, L. García-Díaz, and M. D. C. Romero-Ternero, “Machine Learning to Predict Pre-Eclampsia and Intrauterine Growth Restriction in Pregnant Women,” *Electronics*, vol. 11, no. 19, p. 3240, Oct. 2022, doi: 10.3390/electronics11193240.

[4] X. Hao *et al.*, “Machine Learning Models for Predicting Adverse Pregnancy Outcomes in Pregnant Women with Systemic Lupus Erythematosus,” *Diagnostics*, vol. 13, no. 4, p. 612, Feb. 2023, doi: 10.3390/diagnostics13040612.

[5] A. Raza, H. U. R. Siddiqui, K. Munir, M. Almutairi, F. Rustam, and I. Ashraf, “Ensemble learning-based feature engineering to analyze maternal health during pregnancy and health risk prediction,” *PLoS ONE*, vol. 17, no. 11, p. e0276525, Nov. 2022, doi: 10.1371/journal.pone.0276525.