Volume: 13 Issue: I Month of publication: January 2025

DOI: https://doi.org/10.22214/ijraset.2025.66642

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue I Jan 2025- Available at www.ijraset.com

Analyzing Real-Time Dataset to Understand the Risk Factors of Miscarriage

Sujith Reddy Nalla[1], Ramesh Masuna[2], Adarsh Devajji[3]
ACE Engineering College

Abstract: A miscarriage, also known as a spontaneous loss of pregnancy, occurs naturally and is not intentionally induced. This condition typically takes place within the first trimester (first 20 weeks of gestation), resulting in the loss of the fetus. The impact of a miscarriage extends beyond the physical realm, causing severe emotional and psychological distress for both parents. Understanding and early detection and timely intervention of it will help. These days we have various health sensors [1] that will sense multiple types of health data. using data collected from different devices that would impact miscarriage chances is trained and tested using multiple machine learning algorithms [2] which will help in early detection of miscarriage. A comparative study is performed among various models, and the best model is selected to make accurate predictions.

Keywords: Miscarriage, sensor-data, early detection, machine learning, predictive analysis

I. INTRODUCTION

Pregnancy is a period that typically lasts for nine months, during which a fetus develops in the mother's womb. This time can bring both joy and emotional and physical changes for the expectant mother. Throughout the pregnancy, the baby develops its organs, limbs, and essential functions needed to survive outside the womb. While most pregnancies progress smoothly, approximately 10-15% may end in miscarriage, an unfortunate outcome. A miscarriage affects not only the physical health of the mother leading to potential infections and hormonal imbalances but also has significant emotional consequences, including stress, anxiety, guilt, and depression. These feelings can disrupt personal relationships and diminish one's overall quality of life. Additionally, women may encounter societal challenges after experiencing a loss, which can hinder them from sharing their experiences. For those who endure recurrent miscarriages, concerns about future pregnancies can become even more pronounced.

Miscarriages can occur for various reasons, including chromosomal abnormalities, hormonal imbalances, malnutrition, maternal age, and lifestyle factors. One significant issue is that many women are unaware of their condition during pregnancy. We are addressing this concern by utilizing various sensors and machine learning techniques for predictive analysis, aiming for early detection to prevent negative pregnancy outcomes. By collecting demographic data and lifestyle information such as alcohol consumption and activity levels of pregnant women and training different machine learning models on this data, we strive to improve overall pregnancy outcomes. The importance of this project lies in its continuous collection of real-time data for risk assessment, enabling 24/7 monitoring instead of relying solely on static test results. This approach ensures timely intervention and more accurate predictions, ultimately contributing to better healthcare outcomes for pregnant women. The remainder of the paper is structured to include the following sections: the workflow followed to complete the project, the methodologies employed, a comprehensive literature survey, and an explanation of the proposed system.

A. Data Set

We are utilizing the dataset from the Mendeley repository by Hiba Asri, which consists of 15 attributes and one million records of pregnant women. This dataset includes both activity and lifestyle data, serving as real-time risk factors for predicting the risk of miscarriage. The dataset can be accessed from the following URL:e HIBA ASRI_ Miscarriage Prediction Risk Factors - Mendeley Data

II. LITERATURE REVIEW

[3] San Lazaro Campillo, I., Meaney, S., Sheehan, J. et al. (2018) conducted research with the Pregnancy Loss Research Group at the Irish Centre for Fetal and Neonatal Translational Research, University College Cork (UCC), Ireland, focusing on exploring university students' understanding of the causes and risk factors of miscarriage. Key observations regarding miscarriage risk factors include that fetal chromosomal abnormalities are the most commonly recognized cause, identified by 43% of participants.

Valid risk factors such as advanced maternal age, smoking, alcohol consumption, and maternal medical conditions are observed. However, there are also quite a few misconceptions regarding the risk factors, such as stress, falls or accidents, and drug consumption being causes of miscarriage. Factors like the flu vaccine, flying, and hair dye are mistakenly identified as causes of miscarriage. Notably, stress was more frequently believed to be a risk factor compared to smoking.

[4] Researchers from Christ University's Department of Data Science in Bangalore, India, S. Biswas and S. Shukla (2022), developed machine learning models to predict miscarriage outcomes. They employed a dataset with 10 features and applied three classification algorithms: K-Nearest Neighbors, Logistic Regression, and Random Forests to categorize the data into two groups. The initial dataset contained 1 million records but was highly skewed in terms of age distribution, with 99.7% of individuals being 25 years old. To address this imbalance, the researchers used stratified sampling, which reduced the dataset to 1,775 records. The Random Forest model emerged as the top performer, achieving 97% accuracy. The study had limitations, including the absence of real-time risk factors such as women's daily activities and lifestyle information like alcohol consumption and intoxication levels. While the study attained high accuracy, it only compared three models. Currently, efforts are underway to train and evaluate additional classification models for a more comprehensive analysis.

[5] Hiba Asri, Hajar Mousannif, and Hassan Al Moatassime (2017) focused on leveraging real-time data collected from mobile phones and sensors to predict the chances of miscarriage. The dataset used in this project consists of features like BMI, the number of previous miscarriages, data related to the physical activity of the women, and real-time data such as location. Since the dataset is of large volume, the authors used Apache Spark with Data-bricks for efficient data management and processing, as it uses in-memory processing rather than reading from and writing to disk repeatedly. They made use of the K-Means machine learning algorithm in Apache Spark to segment the data into different risk categories. According to this paper, the dataset is clustered into three groups: miscarriage, no miscarriage, and possible chance of miscarriage. The elbow curve method was used to determine that three clusters were optimal. K-Means predicted 44% of the records as miscarriage, 21% as no miscarriage, and 34% as possible miscarriage. This information is sent to pregnant women via their mobile devices through a simple user interface.

[6] Hiba Asria and Zahi Jarira (2022) collected data from healthcare sensors and mobile phones to predict the risk of miscarriage. Women were equipped with various healthcare sensors to collect real-time risk factors, with data being collected from IoT devices like Raspberry Pi and Arduino. The sensors included temperature sensors, heart rate sensors, alcohol sensors, and acceleration sensors, with Raspberry Pi used to collect and process the data. This dataset includes six additional real-time risk factors compared to the previously proposed dataset. Their further study aims to preprocess the data and use machine learning models on this dataset to create predictive models. We are building our machine learning models on this dataset by borrowing the dataset from this URL.

Based on the reviewed literature, it is evident that significant progress has been made in understanding and predicting miscarriage outcomes using data-driven approaches. Key insights include the importance of recognizing valid risk factors and addressing common misconceptions. Machine learning models, particularly Random Forest, have shown high accuracy in prediction but often lack real-time risk factors such as daily activities and lifestyle data. Utilizing real-time data from mobile phones and sensors has proven effective in segmenting risk categories, while comprehensive datasets including additional real-time features enhance predictive capabilities. However, limitations include the absence of certain risk factors and the need for more diverse and comprehensive models to improve accuracy and reliability in predictions.

A. Architecture

To provide a comprehensive understanding of our system, the following section details the project's architecture. The architecture diagram is included to visually illustrate the system components and their interactions.

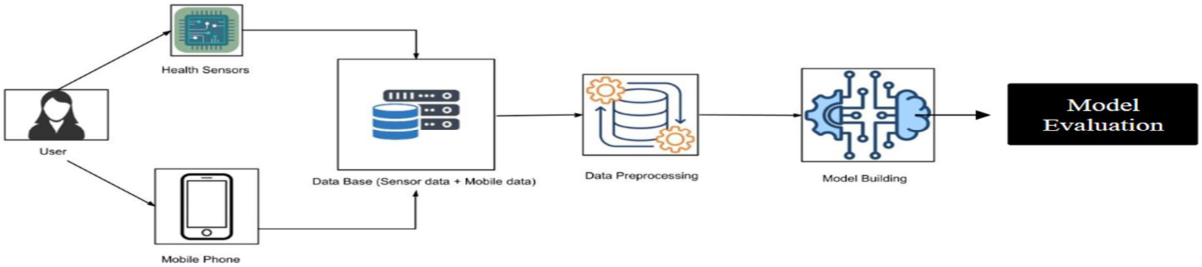

Fig1. System Architecture

III. Proposed Methodology

The proposed methodology employs a range of machine learning algorithms to forecast the probability of miscarriage using a comprehensive dataset gathered from various health sensors and mobile devices. Essential features include BMI, the number of previous miscarriages, physical activity data, and real-time location information. By incorporating diverse machine learning methods such as Logistic Regression[7], Decision Trees[8], Random Forests[9], K-Nearest Neighbors[10], Naive Bayes[11], Neural Networks[12], XGBoost[13], LightGBM[14], CatBoost, and AdaBoost[15], the system seeks to improve predictive accuracy and model reliability.

A thorough comparative analysis is performed across all models, evaluating performance metrics such as accuracy, precision, recall, and F1-score to determine the most effective predictive model. The dataset, enriched with real-time risk factors, provides a solid foundation for predictive analysis. By employing multiple machine learning algorithms and comprehensive evaluation techniques, the proposed system aims to deliver a robust and accurate predictive model for assessing miscarriage risk, ultimately contributing to improved healthcare outcomes for pregnant women.

Using algorithms like **XGBoost** and **LightGBM** known for their efficiency and high performance with large datasets, often achieving state-of-the-art results, significantly enhances the predictive capability of our system. **CatBoost** efficiently handles categorical features, reducing the need for extensive preprocessing. Meanwhile, **AdaBoost** combines multiple weak classifiers to form a strong ensemble model, effectively reducing bias and variance. By leveraging these advanced algorithms, the system aims to provide robust and accurate predictions for assessing miscarriage risk, ultimately contributing to improved healthcare outcomes for pregnant women.

A. Tree-Based Models vs. Standard Machine Learning Models

Compared to algorithms like Logistic Regression, KNN, and Naive Bayes, tree-based models such as Decision Trees, Random Forests, XGBoost, LightGBM, CatBoost, and AdaBoost demonstrate superior performance. This advantage stems from their capacity to manage non-linear relationships and identify intricate feature interactions. These models can directly process categorical variables, minimizing the need for extensive data preparation. Additionally, they show resilience against outliers, which often distort predictions in alternative models. Tree-based approaches offer clear metrics for feature importance, enhancing model interpretability. They efficiently handle large datasets by utilizing parallel processing for faster computations. Unlike Naive Bayes, they are not constrained by naive independence assumptions, and they avoid the dimensionality issues that affect KNN. These attributes make tree-based models particularly effective when dealing with structured data containing complex patterns, resulting in enhanced predictive capabilities.

IV. RESULTS AND CONCLUSIONS

From our current work, we conclude that we have successfully trained and tested various supervised machine learning models after thorough data preprocessing and analysis. We evaluated ten different classification models to predict the chance of miscarriage. After comparing all the models based on the evaluation metrics, the top contenders were AdaBoost and Random Forests. Ultimately, we selected the AdaBoost classifier, as it had a slight edge in precision over Random Forests. AdaBoost achieved an accuracy of 80.81%, a precision of 0.86, a recall of 0.81, an F1-score of 0.80, and a ROC AUC score of 0.93. In contrast, the least performing models, namely Logistic Regression, KNN, and Naive Bayes, yielded accuracies of 61.17%, 63.75%, and 65.39%, respectively.

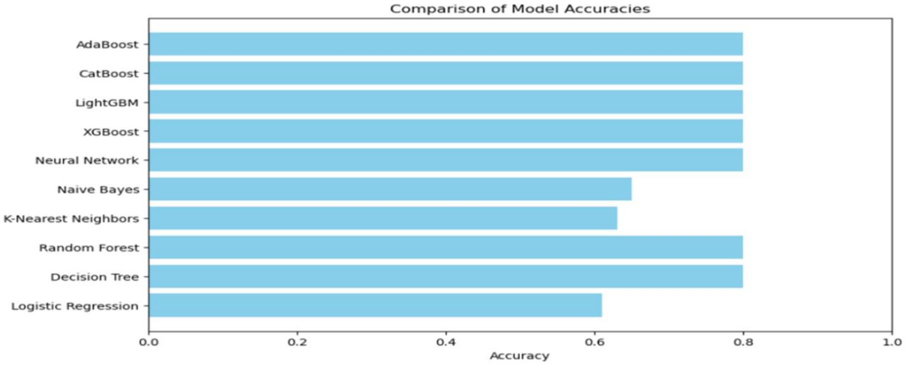

Fig 2. Comparison of Model Accuracies

Model Name	Accuracy(%)	Precision	Recall	F1-Score	ROC AUC
KNN	63.75	0.64	0.64	0.64	0.68
Logistic Regression	61.17	0.61	0.61	0.61	0.64
Decision Tree	80.82	0.81	0.81	0.81	0.81
Naive Bayes	65.39	0.65	0.65	0.65	0.64
Random Forests	80.81	0.81	0.81	0.81	0.93
XGBoost	80.83	0.81	0.81	0.81	0.93
AdaBoost	80.81	0.86	0.81	0.8	0.93
LightGBM	80.83	0.81	0.81	0.81	0.93
CatBoost	80.86	0.81	0.81	0.81	0.93
Neural Networks	80.81	0.86	0.81	0.8	0.93

Fig 4. Models Comparision

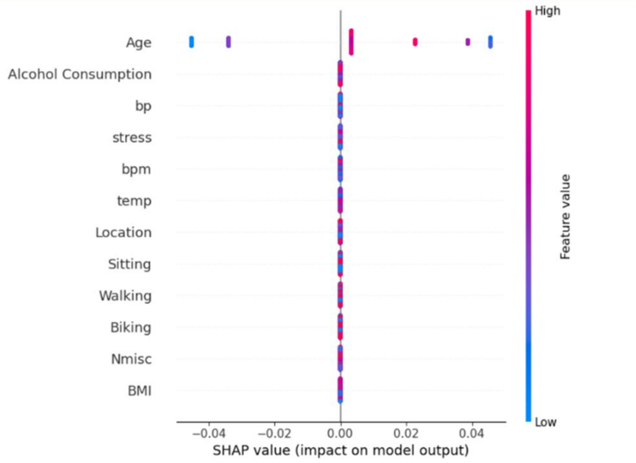

Fig 3. SHAP Summary Plots

Compliance with ethical standards
Disclosure of conflict of interest
No conflict of interest to be disclosed.

REFERENCES

[1] Majumder, S.; Mondal, T.; Deen, M.J. Wearable Sensors for Remote Health Monitoring. Sensors 2017, 17, 130. https://doi.org/10.3390/s17010130

[2] Mahesh, Batta. "Machine learning algorithms-a review." International Journal of Science and Research (IJSR).[Internet] 9.1 (2020): 381-386.

[3] San Lazaro Campillo, I., Meaney, S., Sheehan, J. et al. University students' awareness of causes and risk factors of miscarriage: a cross-sectional study. BMC Women's Health 18, 188 (2018). https://doi.org/10.1186/s12905-018-0682-1

[4] Biswas, S., Shukla, S. (2022). A Miscarriage Prevention System Using Machine Learning Techniques. In: Gupta, D., Khanna, A., Kansal, V., Fortino, G., Hassanien, A.E. (eds) Proceedings of Second Doctoral Symposium on Computational Intelligence. Advances in Intelligent Systems and Computing, vol 1374. Springer, Singapore. https://doi.org/10.1007/978-981-16-3346-1_34

[5] Asri, Hiba & Mousannif, Hajar & Al Moatassime, Hassan. (2017). Real-time Miscarriage Prediction with SPARK. Procedia Computer Science. 113. 423-428. 10.1016/j.procs.2017.08.272.

[6] Asri, Hiba and Zahi Jarir. "Real-time miscarriage prediction: A comprehensive real-world dataset and a new model." FNC/MobiSPC/SEIT (2022).

[7] Odiakaose, Christopher. (2021). A comparative analysis of machine learning algorithms : A case study of a higher institution. 10.13140/RG.2.2.33330.58560.

[8] Kotsiantis, Sotiris & Kanellopoulos, Dimitris & Pintelas, P.. (2006). Data Preprocessing for Supervised Learning. International Journal of Computer Science. 1. 111-117.

[9] Gomez, R., Hafezi, N., Amrani, M. et al. Genetic findings in miscarriages and their relation to the number of previous miscarriages. Arch Gynecol Obstet **303**, 1425–1432 (2021). https://doi.org/10.1007/s00404-020-05859-x

[10] Metwally, M., Ong, K. J., Ledger, W. L., & Li, T. C. (2008). Does high body mass index increase the risk of miscarriage after spontaneous and assisted conception? A meta-analysis of the evidence. Fertility and Sterility, 90(3), 714–726. https://doi.org/10.1016/j.fertnstert.2007.07.1290

[11] LaValley, Michael P. "Logistic Regression." Circulation, vol. 117, no. 18, 2008, pp. 2395–2399. https://doi.org/10.1161/CIRCULATIONAHA.106.682658.

[12] Song, Yan-Yan, and Ying Lu. "Decision tree methods: applications for classification and prediction." Shanghai archives of psychiatry vol. 27,2 (2015): 130-5. doi:10.11919/j.issn.1002-0829.215044

[13] Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. The Stata Journal, 20(1), 3-29. https://doi.org/10.1177/1536867X20909688

[14] Mucherino, A., Papajorgji, P.J., Pardalos, P.M. (2009). k-Nearest Neighbor Classification. In: Data Mining in Agriculture. Springer Optimization and Its Applications, vol 34. Springer, New York, NY. https://doi.org/10.1007/978-0-387-88615-2_4

[15] Leung, K. Ming. "Naive bayesian classifier." Polytechnic University Department of Computer Science/Finance and Risk Engineering 2007 (2007): 123-156.

[16] Kukreja, Harsh, et al. "An introduction to artificial neural network." Int J Adv Res Innov Ideas Educ 1.5 (2016): 27-30.

[17] Torlay, L., Perrone-Bertolotti, M., Thomas, E. et al. Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. Brain Inf. **4**, 159–169 (2017). https://doi.org/10.1007/s40708-017-0065-7

[18] Chengsheng, Tu, Liu Huacheng, and Xu Bing. "AdaBoost typical Algorithm and its application research." MATEC Web of Conferences. Vol. 139. EDP Sciences, 2017.

[19] Angelov, Plamen P., et al. "Explainable artificial intelligence: an analytical review." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 11.5 (2021): e1424.