Grenze International Journal of Engineering and Technology, June Issue

Exploring the Role of Abuse, Substance use and General Health in Pregnancy Risk Prediction using Machine Learning Algorithms

Vijay Sharma[1] and Ravindra Kumar[2]
[1-2]Chandigarh University/ Department of Mathematics, Mohali, India
Email: imvijay900@gmail.com, dr.rkpoonia@gmail.com

***Abstract*—Pregnancy needs continuous health assessment for mothers because it affects maternal wellbeing and baby development. Correct evaluation of pregnancy risks allows doctors to intervene at the right time for proper healthcare distribution. This research examines the performance behavior of today's most advanced machine learning systems. After extensive analysis of multiple approaches, the model selection process was finalized. Its primary task is to classify pregnancy risk levels into three categories: The system generates risk levels of High and Moderate and Low Risk by evaluating the expectant mother's vital signs. Approximately 2,49,654 samples were extracted from the 2016-2022 PRAMS dataset that National Center for Chronic Disease Prevention and Health Promotion (CDC) created. Key input parameters begin with maternal abuse followed by wants to get pregnant then proceed to drinking behavior drug intake and smoking patterns alongside oral health and interactions with doctors about diet ingredients exercise depression anxiety occupational status workplace leave and stress and several preconception and pregnancy vaccine sequences from the mother. Historical patient records containing NIH and WHO defined vital signs and risk assessment data will feed the model which uses Multiple Imputation by Chained Equations (MICE) to handle null values when learning complex input-output connections between variables. An advance version of GBM, Extreme Gradient Boosting shows a highest Accuracy of 91% . This outperformed all other individual classifiers evaluated including Decision Tree (88.65%), Random Forest (82%), Support Vector Machine (71.27%), Gradient Boosting Machine (88.22%), Neural Network (90.43%).**

Through this model healthcare providers maximize resource management by tailoring interventions for pregnant patients. This novel ML system produces accurate pregnancy risk forecasts which can enhance both maternal and fetal healthcare by enabling quick interventions while streamlining personalized care approaches for better worker quality.

***Index Terms*— Pregnancy Risk Prediction, Extreme Gradient Boosting (XGBoost), PRAMS Dataset Analysis, Predictive Modeling in Healthcare, Personalized HealthCare Worker.**

I. INTRODUCTION

Public health officials emphasize maternal health status throughout pregnancy because pregnancy-related complications generate different health effects on expectant mothers and newborns. Pregnancy-related risk assessment alongside continuous monitoring serves as an essential foundation to stop negative birth-related

Grenze ID: 01.GIJET.11.2.570

effects and support general fetal and gestational health throughout the stages of pregnancy and birth. Several obstacles exist in building reliable maternal health risk assessment technology because pregnancy complications stem from multiple interacting behavioral, clinical and environmental elements.

The research Explore the Role of Abuse, Substances used, General Health in Pregnancy Risk Prediction through the integration of vast data sets from the Pregnancy Risk Assessment Monitoring System (PRAMS) which is administered by the National Center for Chronic Disease Prevention and Health Promotion (CDC). Through its data collection from maternal experiences across prenatal and postpartum stages PRAMS offers researchers extensive insights into pregnancy risk-determining factors. The research gathers PRAMS datasets from 2016 to 2022 to build a unified risk assessment framework through more than 42 different combination sets of health condition and variable datasets. All datasets underwent preprocessing steps which selected time-defined pregnancy variables from before pregnancy until delivery while normalizing medical condition attributes across columns. A complete dataset was created with 249,556 samples by uniting various features together. Analyzing missing data became the key initial step due to its prevalence in major health databases. Multiple Imputation by Chained Equations (MICE) provided an effective method for handling missing values thus maintaining dataset reliability for further analytic processes. A medical risk scoring system evaluated sample pregnancy risk through predetermined risk factors that affect maternal as well as fetal well-being. The risk levels were categorized into three groups: low, moderate, and high. The proposed Pregnancy Risk Level target variable served as the foundation to develop ML models that forecast risks and detect high-risk maternal situations. A set of classification algorithms was used for model performance evaluation including Decision Tree and XGBoost and Random Forest. XGBoost achieved the greatest accuracy rate by delivering a 91% performance result. The model demonstrates robust functionality with its capacity to process advanced data structures as well as uneven risk category distributions. The design of the pregnancy risk assessment system has developed a data-based methodology that helps to monitor and identify high-risk pregnancies. The research adds value to modern maternal health improvement endeavours through its capability to initiate proactive interventions based on personalized care strategies.

II. Literature Review

Assessing potential pregnancy risks represents a fundamental aspect when parent counsellors work with healthcare providers to identify potential complications mechanisms and prevent adverse effects for mother and baby throughout maternal care.

Healthcare providers must determine pregnancy risks because it helps prevent medical complications. Maternal and fetal health assessments have benefited substantially from developments in data science along with monitoring technology.

[1] delivered an online risk assessment system using combined blood pressure analysis and glucose measurement and pulse rate data which reached an accuracy rate of 81.5%. This methodology using Machine Learning technology helps medical facilities allocate their resources more effectively.

Maternal heart disease risks increase dramatically when cardiovascular changes occur during pregnancy but the risks are especially significant for women with CHD. The authors stress that individualized risk scoring together with interprofessional heart teams should be combined with regular check-ups to optimize pregnancy results [2] explains.

Radiological advances in wireless body networking (WBN) technologies have improved maternal risk assessment throughout pregnancy. The development of [3] introduced fetal monitoring technology based on ultra-wideband transducers which combines dependable remote healthcare capabilities with superior signal performance capabilities.

Neuro-fuzzy systems along with fuzzy logic represent powerful methods for assessing fetal health conditions. Researchers [4] implemented a neuro-fuzzy system driven by Doppler ultrasound to classify fetal health levels into good and suspect and alarming categories for clinical decision support.

PRAMS provides data that enables research focused on maternal healthcare. The evaluation of postpartum depression risk using Random Forest and TabNet and SVM models demonstrated that TabNet provided the best Area Under the Curve result (0.7779) along with SVM's optimal F1 score by researchers [5]

Analysing Medicaid coverage under the ACA exposed postpartum healthcare access deficiencies despite offering prenatal benefits according to [6]. The research backs policy measures that will expand postpartum healthcare access longer than 60 days.

The integrated system of IOT [7] uses mobile applications together with wearable technology to provide continuous maternal health surveillance. The analysis documented its successful implementation for proactive

maternal-fetal healthcare interventions.
A microfluidic system developed by [8] examined nanoparticle impact on fetal well-being while emphasizing the requirement of safe nanomedicine in maternal healthcare.
Medal honoured the need for multidisciplinary cardiovascular risk assessments performed on pregnant women with heart conditions. Their [9]integrated framework utilizes imaging tests together with biomarker evaluations for both early risk identification and personalized healthcare delivery.
Researchers [10] explore blood pressure monitoring serves as a fundamental tool to control pregnancy complications that include preeclampsia. This review highlighted conventional method weaknesses, which led to proposals for wearable technology enhancements.
Motherhood behaviour shows changes because of legalized cannabis laws. The legalization of marijuana corresponded to increased consumption before and after pregnancy according to researcher [11] without sufficient evidence of fetal health implications.
Team of [12] investigated how ensemble learning methods working with feature engineering techniques could optimize pregnancy risk prediction benchmarks. The approach provides early warning signals about health deterioration which leads to better real-time maternal care assessments and rapid intervention efforts.

III. METHODOLOGY

A. Overview

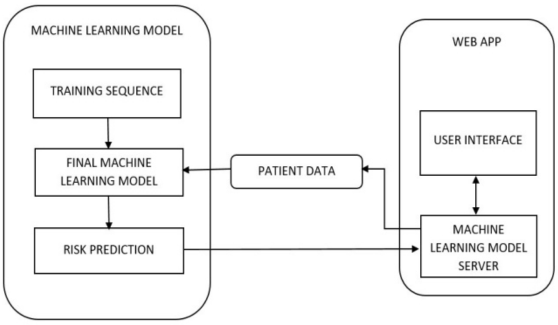

Fig. 1: Architecture diagram

The diagram in Figure1. illustrates how data moves through the system according to the project methodology. The system comprises the following components:

Machine Learning Model
Training Sequence: The development cycle requires this processing phase because it shapes the final ML model which performs predictions. A CSV file dataset serves as the starting point for a process which employs multiple statistical methods and machine learning techniques during extensive data preprocessing. Pregnancy Risk is created as a target variable through assessment of different medical conditions as healthcare experts refer to the NHI and WHO standards. The project utilizes four ensemble classifiers: Gradient Boosting Classifier, Decision Tree, Random Forest and Extreme Gradient Boosting.
Final Machine Learning Model
XGBoost proved to be the superior model for preprocessing data through its extensive training phase leading to highest accuracy rates. The robust prediction capabilities alongside intricate data relationship management qualities make XGBoost the preferred choice for our Pregnancy Risk Assessment Monitoring System.
Risk Prediction
The conclusion model demonstrates superior performance in maternal risk assessment when gauging possible complications that involve substance abuse and mental health problems. Relevant predictions emerge from the relationships found among training data input variables which the model extracted during its training phase.

Patient Data:
Survey data consists of observations about Abuse by husband, Alcohol consumption and smoking patterns and Consume Drug in any form behaviors, wants to get Pregnant before conception, Diet & Exercise habits in addition to Counselling with HCW and Depression & Anxiety issues and including work during pregnancy as well as Oral Health status and how often Patients get help and PreMShots vaccination and TDap vaccination during pregnancy data greatly influence the accuracy of risk predictions.

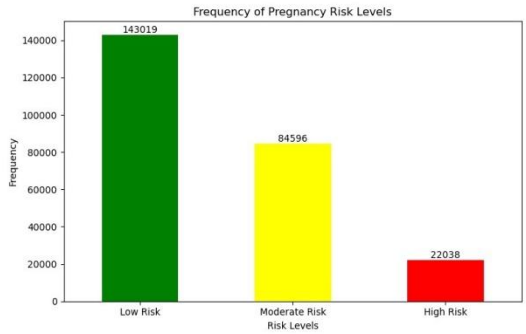

Fig. 2: Frequency Distribution of Each Risk Level

Web App

User Interface

The front-end section of the web-based application provides users with their main point of engagement. The application contains a system that allows users to access risk assessments after they input patient information.

Machine Learning Model Server

As the web application's backend element the model server maintains the trained ML model for use in risk assessments. The system receives user interface information to process it through the ML model which produces predictions concerning risks before displaying results on the user interface. The server manages both model deployment operations along with prediction request processing.

B. Machine Leaning Techniques used:

Multiple Imputation by Chained Equations

The advanced statistical method known as MICE substitutes missing values from data sources through estimated reasonable values constructed from standard observed dataset inputs. Multiple chained repetitions of this imputation method produce several complete datasets which assign distinct estimated values independently representing different levels of uncertainty from the imputation process.

Decision Tree

Decision tree brings supervised learning through a method that performs both classification and regression outputs. Through the decision tree algorithm data partitioning identifies features that yield optimal information value based on the Gini index. Decision Trees operate through nodes that present decisions followed by outcome branches that lead to final prediction results exposed at the leaves. This algorithm allows users to understand it effectively as it operates equally well with both categorical and numerical data types. Overfitting remains a problem for this algorithm since maximizing tree complexity reduces stability and dramatically affects group optimization using minimal dataset alterations.

Random Forest

The ensemble learning method Random Forest engages in building numerous decision trees through random selection of features combined with bagging (Bootstrap Aggregating) data subsets. Decision trees become trained with randomly selected data parts and nodes choose features randomly during their operation through this method. All decision trees become operational after their creation results in model predictions. In random forest solutions the majority voting mechanism selects the tree prediction with the most votes for classification problems. Regression prediction outputs use the average values from each tree model in the ensemble.

By using many decision trees in its model Random Forest maintains flexible prediction abilities which generate accurate predictions across different tasks.

Gradient Boosting machine

Gradient Boosting creates advanced ensemble techniques through which decision trees are constructed in sequential order after each preceding tree finishes its development. The insertion of new trees within the ensemble helps to establish refined predictions for any remaining prediction errors that earlier trees in the model did not address. A learning rate that works jointly with loss function minimization supports effective processing of wideranging data connections within the system. Gradient Boosting delivers robust performance although its demands substantial computing power and includes susceptibility to overfitting resulting from imperfect hyperparameter configuration covering learning rate control and tree depth optimization. THE iterative nature of the system makes it capable of handling complex dataset configurations successfully. Model enhancement along with faster executions depends on dedicated hyperparameter optimization which leads to trustworthy prediction results.

Extreme Gradient Boosting

The updated version Extreme Gradient Boosting improves both the speed and performance of Gradient Boosting algorithms. This advance version can perform both classification and regression tasks. The performance enhancement framework of XGB features efficient memory techniques together with parallel computation and both L1 and L2 regularization methods. XGBoost builds a sequence of connected trees which uses both tree shrinking and subsampling and scaling features to minimize overfitting and reduce loss errors. The level of precision and reliability in ML technology reaches its peak with XGBoost which effectively processes databases that include any missing point values. To achieve optimal performance XGBoost requires fine-tuning its exact optimal hyperparameter settings.

Neural Network

A Neural Network served as the classification technique for this investigation. The ReLU-activated multilayered architecture trained through Adam optimization served as the basis for the implementation. Training stability was improved through normalization applied to the data followed by the use of backpropagation to reduce categorical cross-entropy loss. The analysis implemented dropout regularization as a measure to reduce overfitting. The model assessment included accuracy scores and precision levels and recall functionality together with the F1-score metric. Additionally the process of hyperparameter optimization used grid search methodology to determine the learning rate parameter values as well as batch size parameters and hidden layer quantities.

Support Vector Machine

A SVM ran parallel to other classification methods on the data set. RBF and polynomial kernels allowed the model to find an optimal hyperplane which classified data in spite of potential non-linear separability. Standardization was implemented across the features to maintain distributed contribution levels. The regularization parameter C and kernel coefficient gamma were optimized using Grid search together with cross-validation parameters. A confusion matrix with accompanying accuracy, precision and recall and F1-score evaluation established the effectiveness of the implemented model.

IV. Result And Discussion

The proposed Pregnancy Risk Prediction System was effective in classifying pregnancy risk levels in a dataset of maternal health factors. The dataset was prepared for machine learning model after missing values were handle with MICE technique.

This initiative is highly promising as it can help clinicians identify pregnancy risk factors much earlier in the game. This is achieved by segmenting a pregnant woman's health metrics (E.g., Abuse by husband, family or other, Alcohol consumption, smoking, Consume Drug in any form, want to get Pregnant, Diet & exercise, Counselling with HCW, Depression & anxiety, work during pregnancy, Oral health problems, get help, PreMShots, and TDap vaccination during pregnancy) is classified out as low, moderate, or high. Such categorization allows healthcare providers to individualize prenatal care and interventions.

For categories such as medical diagnosis, greater precision (e.g., >90%) is frequently necessitated. This model powered by the data makes predictors of 91% accuracy in detecting patterns and relationships between those health metrics and levels of pregnancy risk. Nonetheless this accuracy can be improved with constant monitoring and model tuning. However this needs to be put into practice in close cooperation with healthcare professionals that will accompany the deployment process to maximally align with mothers' convenience.

Overall, this project highlights the use of Machine Learning in maternal healthcare and supports healthcare providers by offering helpful risk assessment tools to improve the care and treatment of pregnant women.

The key results for the various classification models evaluated are as follows:

TABLE I: ACCURACY COMPARISON OF DIFFERENT ALGORITHMS

S.No	Algorithm	Accuracy(%)
1	Decision Tree	88.65
2	SVM	71.27
3	Random Forest	82
4	GBM	88.22
5	Random Forest	91
6	Neural Network	90.43

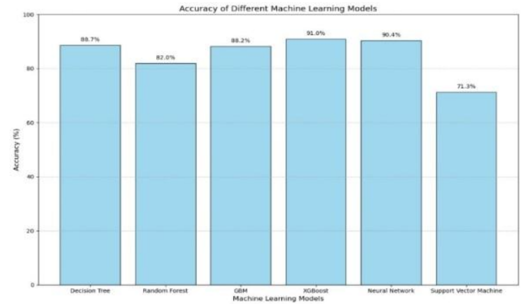

Fig. 3: graphical Representation of Accuracy of Different Classifier Algorithms.

Fig. 3. compares the accuracy of various classifier algorithms used for maternity risk prediction. It can be seen that the model, Extreme Gradient Boosting, shows the highest accuracy.

A. Decision Tree

The Decision Tree performed moderately 88.65% but displayed signs of overfitting since the simplicity of the algorithm restricted its capacity to generalize across different levels of risk.

B. Random Forest Tree

Random Forest is an ensemble model and hence it avoids the overfitting and gain stability over Decision Tree, its accuracy is like 82%.

C. Gradient Boosting Machine

While this still kept the tree as the core of the model (a weak learner), GBM worked on a better bias-variance trade-off of flexibility vs accuracy reducing the residual errors iteratively with each new tree added: [achieved: 88.22%].

D. Neural Network

Captured Non-Linear Relationships, but needed Hyperparameter Tuning & Large Computational Power MSE 90.43%

E. Support Vector Machine

TThe Support Vector Machine (SVM) model generated 71.2% accuracy when used to forecast pregnancy risks through analysis of abuse conditions and substance use habits along with smoking patterns and dimensions of healthcare support and vaccination rates and dietary patterns. SVM presents itself as a suitable solution for detecting high-risk pregnancies yet further algorithm optimization together with alternate ML techniques might boost its predictive power. To enhance model reliability and performance future projects should explore both ensemble learning methods alongside feature selection techniques.

F. Extreme Gradient Boosting

The XGBoost model achieved 91% accuracy as its highest performing metric. The data review system demonstrated the ability to address missing values while utilizing L1 and L2 penalties for regularization and taking advantage of parallel processing to speed up operations.

The concluded XGBoost model performed well through its precision, recall and F1 scores across the full range of pregnancy risk categories. High-risk pregnancy categorization demonstrated successful results through this model because such identifications need immediate medical care. The strong model performance indicates its potential value as an approach to reliably assess pregnancy risk.

The confusion matrices of each model reflect their accuracy rate by displaying actual label segregation versus predicted label outcomes.

V. CONCLUSION

The Pregnancy Risk Assessment and Monitoring System established through this research illustrates machine learning's potential to forecast varied maternal health components for pregnancy risk assessment levels. A combination of extensive PRAMS database information alongside MICE pre-processing and multiple ML classifiers produced this research's optimal predictive method. Through the use of XGBoost and Gradient Boosting with Random Forest the classification system reached 91% accuracy which established its reliability for pregnancy risk assessment.

Artificial Intelligence techniques now show great promise to help maternal healthcare by detecting early pregnancy risks to support swift medical assessments. Through a web application that offers user-friendly interfaces healthcare providers benefit from potent assessments of risk levels alongside improved resource allocation capabilities. This research demonstrates how ML systems enhance individualized prenatal care while using patient data to deliver better outcomes for both mothers and babies.

The next phase of work must focus on model optimization using real-time health data while increasing population coverage of the model alongside improvements in its existing data representation biases. The successful implementation of this maternal healthcare framework depends on joint efforts between healthcare professionals who will perform clinical validation and system implementation.

The research demonstrates artificial intelligence's power to transform public health by developing precise cost-effective pregnancy risk assessment solutions for global maternal care improvement.

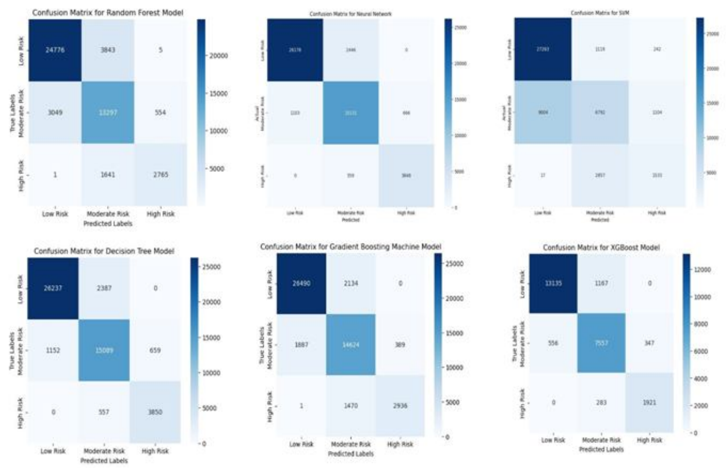

Fig. 4: Comparison of Confusion Matrix of Different Algorithms

FUTURE ASPECTS

The proposed Pregnancy Risk Assessment and Monitoring System demonstrates the potential of ML in maternal healthcare, but several areas can be explored further to enhance its effectiveness and applicability

1) Integration with Real-Time Health Monitoring Systems:Subsequent versions of this model should include wearables with mobile health tools for getting time-sensitive physiological and behavioral information from pregnant women. Watching maternal health indicators consisting of heart rate and blood pressure and glucose levels through continuous monitoring builds improved risk assessments and permits transparent medical care.

2) Expansion of Dataset and Multilingual Data Inclusion: The present research depends on PRAMS dataset information yet it requires additional maternal health data from around the world to improve the model's overall applicability. The integration of multilingual data brings benefits through simpler access for diverse maternal populations during risk management assessments.

3) Personalized Risk Prediction and Recommendation System: Future iterations of the system will deliver personalized interventions through deep learning algorithms alongside natural language processing(NLP) models to detect risks and match solutions with individual health information.

4) Clinical Validation and Deployment in Healthcare Systems: Healthcare professionals together with medical institutions need to validate the system along with each other. Hospital and maternity clinic pilot tests will demonstrate the system's true operational capabilities in supporting obstetricians and gynecologists.

REFERENCES

[1] M. F. Ukrit, R. B. Jeyavathana, A. L. Rani, and V. Chandana, "Maternal health risk prediction with machine learning methods," in 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE). IEEE, 2024, pp. 1–9.

[2] I.M. van Hagen and J. W. Roos-Hesselink, "Pregnancy in congenital heart disease: risk prediction and counselling," Heart, vol. 106, no. 23, pp. 1853–1861, 2020.

[3] E. Palantei, I. S. Areni, A. A. Pratiwi, and A. Amir, "Improvement of uwb patch transducer properties applicable for

fetal monitoring system,” in 2015 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob). IEEE, 2015, pp. 192–195.

[4] F. Gurgen, N. Guler, and F. Varoll, “Antenatal fetal risk assessment using a neurofuzzy technique,” IEEE engineering in medicine and biology magazine, vol. 20, no. 6, pp. 165–169, 2001.

[5] A. Paul, S. D. Pragada, D. N. Murthy, M. L. J. Shruthi, and S. Gurugopinath, “Performance comparison of machine learning techniques for early detection of postpartum depression using prams dataset,” in 2023 IEEE 15th International Conference on Computational Intelligence and Communication Networks (CICN). IEEE, 2023, pp. 310–315.

[6] E. M. Johnston, S. McMorrow, C. Alvarez Caraveo, and L. Dubay, “Postaca, more than one-third of women with prenatal medicaid remained uninsured before or after pregnancy: Study examines insurance coverage and access to care before, during, and after pregnancy for women with prenatal medicaid coverage,” Health Affairs, vol. 40, no. 4, pp. 571–578, 2021.

[7] F. Sarhaddi, I. Azimi, S. Labbaf, H. Niela-Vilen, N. Dutt, A. Axelin,´ and A. M. Rahmani, “Long-term iot-based maternal monitoring: system design and evaluation,” Sensors, vol. 21, no. 7, p. 2281, 2021.

[8] S. Shojaei, M. S. Ali, M. Suresh, T. Upreti, V. Mogourian, M. Helewa, and H. I. Labouta, “Dynamic placenta-on-a-chip model for fetal risk assessment of nanoparticles intended to treat pregnancy-associated diseases,” Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, vol. 1867, no. 7, p. 166131, 2021.

[9] G. Sharma, W. Ying, and C. K. Silversides, “The importance of cardiovascular risk assessment and pregnancy heart team in the management of cardiovascular disease in pregnancy,” Cardiology clinics, vol. 39, no. 1, pp. 7–19, 2021.

[10] A. Hurrell, L. Webster, L. C. Chappell, and A. H. Shennan, “The assessment of blood pressure in pregnant women: pitfalls and novel approaches,” American journal of obstetrics and gynecology, vol. 226, no. 2, pp. S804–S818, 2022.

[11] K. R. Skelton, A. A. Hecht, and S. E. Benjamin-Neelon, “Association of recreational cannabis legalization with maternal cannabis use in the preconception, prenatal, and postpartum periods,” JAMA network open, vol. 4, no. 2, pp. e210138–e210138, 2021.

[12] 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, 2022.