Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 20s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Data Wrangle and Kurtosis Matching Regression Based Machine Learning for Maternal Health Risk Prediction

C MidhunaMurali[1], Dr.P Senthil Vadivu[2]

[1]Research Scholar, Department of Computer Science, Hindusthan College of Arts & Science College, Coimbatore, India

Email ID: midhumurali1903@gmail.com

[2]Professor & Head, Department of Computer Applications, Hindusthan College of Arts & Science College, Coimbatore, India

Email ID: senthilvadivu.p@hicas.ac.in

Cite this paper as: C MidhunaMurali, Dr.P Senthil Vadivu, (2025) Data Wrangle and Kurtosis Matching Regression Based Machine Learning for Maternal Health Risk Prediction. *Journal of Neonatal Surgery*, 14 (20s), 255-263.

ABSTRACT

Prediction of health disease employing ML algorithm system using predicting a patient's illness based on observed symptoms. Proposed Scale Normalized Data Wrangle and Kurtosis Matching Regression (SNDW-KMR) is introduced for maternal health risk prediction. Initially, pregnancy risk factor dataset is considered as input. Robust Scaled Normalization Process and Box Plot Data Wrangling are performed for normalization and outlier detection. Features necessitated for pregnancy health risk prediction is selected employing Weight Jarque–Bera Kurtosis Matching Regressive Feature Selection algorithm. Experimental evaluation conducted with several metrics.

Keywords: *Machine Learning, Maternal Health Risk Prediction, Scale Normalization, Data Wrangling, Weight Jarque–Bera, Kurtosis Matching Regression*

1. INTRODUCTION

Maternal health risk prediction is critical feature as public health is concerned. An advanced machine learning technique was designed in [1] to forecast Gestational diabetes mellitus (GDM) risk in pregnant women. But error and outlier detection rate involved risk in pregnant women was not focused. A novel parameter optimization method using MRA-optimized SVM was proposed in [2] to improve maternal health risk prediction. But F-measure and false positive rate involved in maternal health risk prediction was not discussed. Weight Jarque–Bera Kurtosis Matching Regressive a clinical pregnancy prediction model was introduced in [3] for implementing ML. Bayesian approach was designed in [4] to predict weight gain while dealing with limited data availability. Ensemble ML classifier was introduced in [5] for features extraction by classification. End-to-end overview of FetalAI's development process was carried out in [6] to perform efficient user interactions. Prediction methods were developed in [7] and [8] to predict Period Score during delivery period. In [9], fundamentals of ML different prediction methods were designed. Prevailing management techniques consists of early identification and initiation of risk circumventing interventions eased by a rules-based checklist [10]. In [11] ML were applied with higher accuracy. Yet another method employing ensemble of ML technique was designed in [12]. A systematic review of ML based decision support system for different aspect of maternal health care prediction was investigated in [13]. An overview on evolution of ML techniques was designed in [14]. Two novel ensemble methods for miscarriage prediction was presented in [15] with main focus on error rate.

1.1 Contributions of the work

- To improve maternal health risk prediction accuracy, SNDW-KMR model is developed
- To reduce error rate, Robust Scaled Normalized and Box Plot Data Wrangling is used.
- To select relevant features, Weight Jarque–Bera Kurtosis Matching Regressive Feature Selection algorithm denoising model is applied
- To conduct proposed SNDW-KMR methods are compared to existing [1] and [2].

Related Works

In [16], ML technique was applied for predicting risk on basis of immune abnormalities. In [17], correlation based feature selection employing ML technique was proposed that based on high correlated features provided. A predicting maternal health risk employing sophisticated ML techniques was investigated in [18]. MLP-NN classifier was proposed in [19] of predicting depression risk and anxiety in pregnant women. Yet another clinical-setting model to rule out unfavorable maternal aftermaths in women employing novel ML method was presented in [20]. Several ML algorithms were proposed in [21] for predicting moderate-to-severe depression. In [22], a plethora of ML algorithms in maternal risk level prediction employing global maternal mortality dataset from Oman was designed. A hybrid method combining IoT and big data for real time maternal risk prediction was presented in [23]. In [24], three distinct ML algorithms and decision tree regression were designed to predicting maternal health risk prediction. Yet another real time maternal health prediction method employing SVM and ANN was presented in [25]. A literature survey on ML based box methods were proposed in [26] for both ensuring pregnancy care. In [27] a method to predict risk associated with maternity using principal component analysis and stacked ensemble voting classifier was presented. By employing these two mechanisms resulted [28] in improvement of precision. Yet another method employing deep hybrid method, artificial neural network and random forest algorithm was proposed in [29]. A robust data scaling algorithm was designed in [30] to improving accuracy of detection rate.

2. PROPOSAL METHODOLOGY

ML method for maternal health risk prediction, called SNDW-KMR is choosing the relevant features with high accuracy.

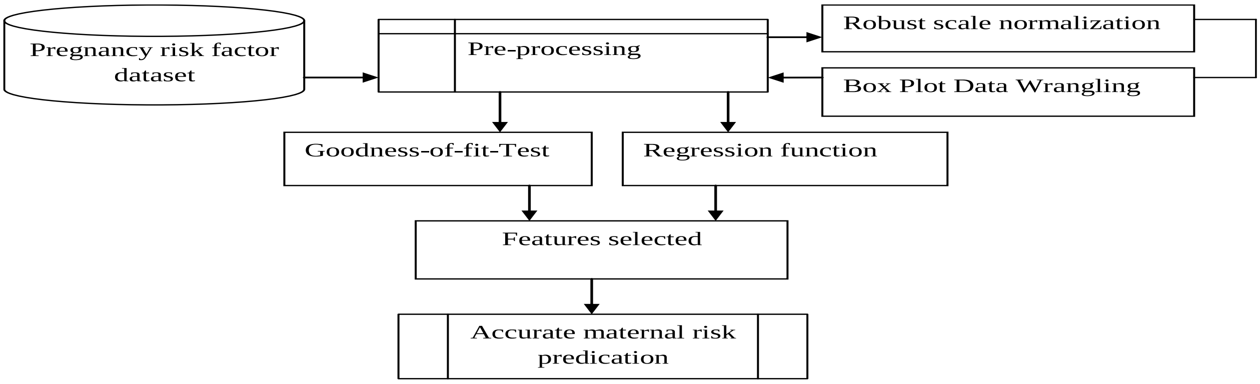

Figure 1 Structure of (SNDW-KMR) method

In figure1, the proposed method performs two distinct processes, data pre-processing and feature selection.

2.1 Dataset description

The pregnancy risk factor dataset employed in our work for maternity health risk prediction consists of eleven features and 6104 sample instances. the dataset including 11 Features such as , Patient ID, Name, Age, Body temperature , Heart rate, Systolic blood pressure, Diastolic blood pressure , BMI, Blood glucose, Blood glucose, Outcome.

2.2 Robust Scaled Normalized and Box Plot Data Wrangling-based Preprocessing model

In this work a pre-processing model combining Robust Scaled Normalized and Box Plot Data Wrangling is designed

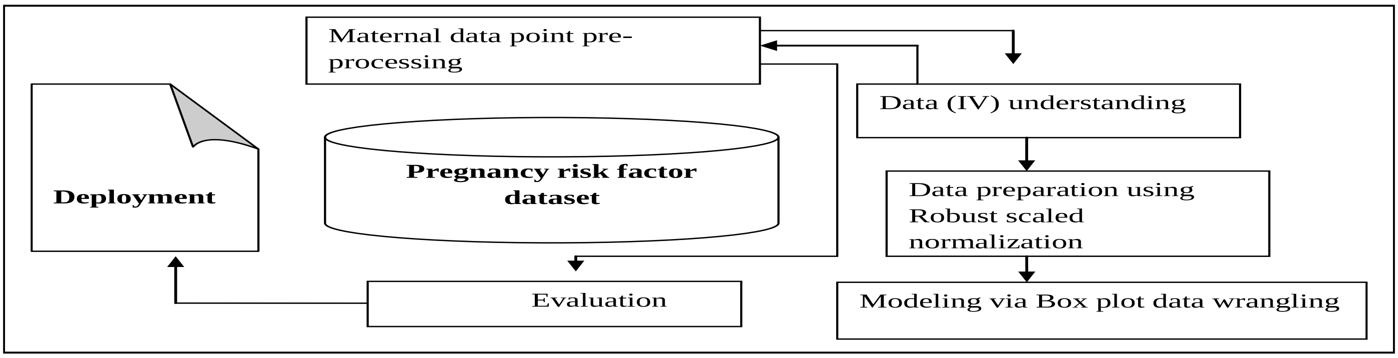

Figure 2 Structure of Robust Scaled Normalized and Box Plot Data Wrangling model

In figure 2, maternal data point pre-processing, input vector formulated for raw pregnancy risk factor dataset. Data preparation using Robust Scaled Normalization Process to scale data points in specific range. Box Plot Data Wrangling easily detects outliers. Normalized and outlier remove features input vector are ready for evaluation and deployment. Consider pregnancy risk factor dataset 'DS' as input with 'm' features 'F' and 'n' samples 'S' formulated as input vector.

$$IV = \begin{bmatrix} S_1F_1 & S_1F_2 & \dots & S_1F_m \\ S_2F_1 & S_2F_2 & \dots & S_2F_m \\ \dots & \dots & \dots & \dots \\ S_nF_1 & S_nF_2 & \dots & S_nF_m \end{bmatrix} \quad (1)$$

In (1), input vector 'IV' is described. Robust Scaled Normalization function is applied next for employing median and IQR to robust outliers. It scales feature values present in input vector utilizing IQR. Here Robust Scaled Normalization function for 'IV' as follows,

$$NIV = \frac{IV - Q_2(IV)}{Q_3(IV) - Q_1(IV)} \quad (2)$$

In (2), '$Q_2(IV)$', '$Q_3(IV)$' and '$Q_1(IV)$' specifies three quartiles '$25th$', '$50th$' and '$75th$' quartile of feature values present in input vector. With obtained 'NIV' results, wrangling function is applied for processing data and detect outliers. By performing munging or sorting, data is processed for resultant feature values. Second quartile or median results are obtained.

$$Q_2[IV] = \begin{cases} NIV\left[\frac{n+1}{2}\right], if\ n\ is\ odd \\ \frac{NIV\left[\frac{n}{2}\right] + NIV\left[\frac{n}{2}+1\right]}{2}, if\ n\ i\ even \end{cases} \quad (3)$$

From (3), middle maternal data point value is selected for odd number of observations and mean of two maternal data points are selected in case of even number of observations. First quartiles '$Q_1[IV]$' and third quartiles '$Q_3[IV]$' are obtained by dividing into two.

$$Q_1[IV] = \sum Med(lower\ region)/2 \quad (4)$$

$$Q_3[IV] = \sum Med(upper\ region)/2 \quad (5)$$

In (4) and (5) first quartiles '$Q_1[IV]$' and third quartiles '$Q_3[IV]$' results are arrived at averaging of median of lower region '$Med(lower\ region)/2$' and averaging of median of upper region '$Med(upper\ region)/2$'. Inter quartile region is obtained and range of box plot is,

$$IQR = Q_3[IV] - Q_1[IV] \quad (6)$$

From (6), inter quartile region 'IQR' is evaluated by subtracting '$Q_1[IV]$' from '$Q_3[IV]$'. Next, range of box plot is obtained as,

$$Range: Q_1[IV] - 1.5 * IQR\ [lower\ part\ of\ range];\ Q_3[IV] + 1.5 * IQR[upper\ part\ of\ range \quad (7)$$

In (7), by maternal data points going outside the range then it's an outlier and vice versa.

Input: Dataset 'DS', Samples '$S = \{S_1, S_2, \dots, S_n\}$', Features '$F = \{F_1, F_2, \dots, F_m\}$'
Output: outlier-eliminated pre-processed data samples 'PD'
Step 1: **Initialize** '$n = 6104$', '$m = 11$' Step 2: **Begin** Step 3: **For** each Dataset 'DS' with Samples 'S' and Features 'F' Step 4: Formulate input vector according to (1) Step 5: Compute Robust Scaled Normalization function for the input vector 'IV' according to (2) Step 6: Formulate positioning with smallest numbers left-positioned and largest numbers right-positioned Step 7: Evaluate second quartile or median results according to (3) Step 8: Evaluate first quartile and third quartile according to (4) and (5) Step 9: Compute inter quartile region according to (6) Step 10: If '$Range > NIV$' Step 11: **Then** outliers and maternal data points removed Step 12: **End if** Step 13: **If** '$Range \leq NIV$' Step 14: **Then** no outliers and maternal data points are retained for further processing Step 15: **Return** pre-processed data samples 'PD' Step 16: **End if** Step 17: **End for** Step 18: **End**

Algorithm 1 Robust Scaled Normalized and Box Plot Data Wrangling

2.3 Weight Jarque–Bera Kurtosis Matching Regressive Feature Selection model

Weight Jarque–Bera Kurtosis Matching Regressive Feature Selection algorithm is designed to select the relevant features for maternal health risk prediction.

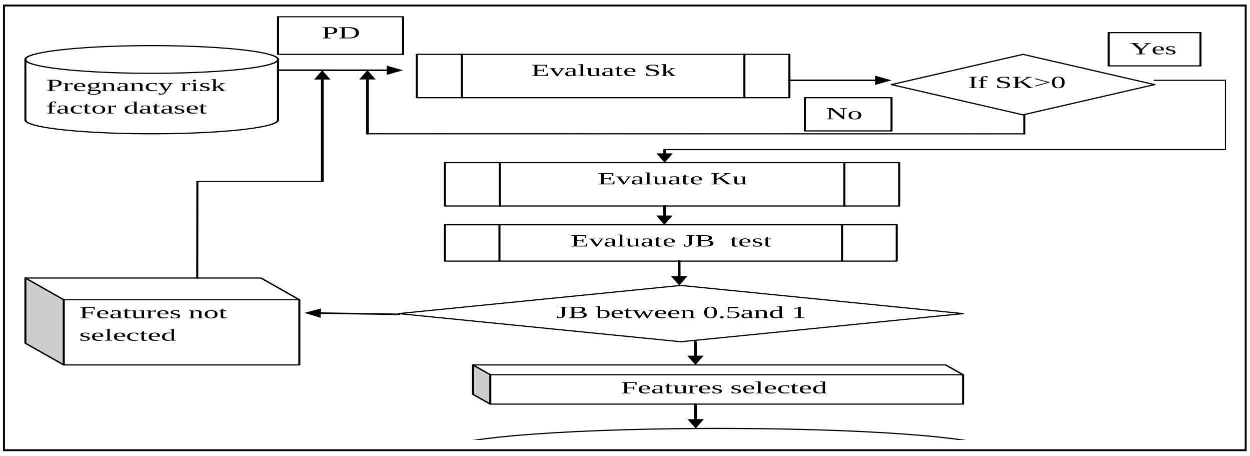

Figure 3 Flow diagram of Weight Jarque–Bera Kurtosis Matching Regressive Feature Selection

Figure 3, pre-processed samples and features are considered to perform Weight Jarque–Bera Kurtosis Matching Regressive Feature Selection model. With preprocessed data samples, skewness 'Sk' and sample kurtosis 'Ku' for corresponding 'm' set of feature is as follows.

$$Sk = \frac{\frac{1}{m}\sum_{i=1}^{m}(F_i-\bar{F})^3}{\left(\frac{1}{m}\sum_{i=1}^{m}(F_i-\bar{F})^2\right)^{3/2}} \quad (8)$$

$$Ku = \frac{\frac{1}{m}\sum_{i=1}^{m}(F_i-\bar{F})^4}{\left(\frac{1}{m}\sum_{i=1}^{m}(F_i-\bar{F})^2\right)^{2}} \quad (9)$$

In (8) and (9), 'Sk' and 'Ku', with respect to features are evaluated. The value 'Sk' is greater than '0' denotes right-skewed and 'Sk' less than '0' denotes left-skewed. Then, Jarque–Bera Test is examined with its 'Sk' and kurtosis is stated as.

$$JB = \frac{l}{6}\left(Sk^2 + \frac{1}{4}(Ku-3)^2\right) \quad (10)$$

With Jarque–Bera Test, Kurtosis matching goal is find weights for each observation in regression that, result of residuals with kurtosis value closer to normal distribution. Finally, relevant features are considered for maternal healthcare prediction.

Input: Dataset 'DS', Samples '$S = \{S_1, S_2, \dots, S_n\}$', Features '$F = \{F_1, F_2, \dots, F_m\}$'
Output: accurate and relevant features selected
Step1: **Initialize** '$n = 6104$', '$m = 11$', preprocessed data samples '$PD = \{PD_1, PD_2, \dots, PD_l\}, l \le n$' Step 2: **Initialize** '$normal\ distribution = 3$' Step 3: **Begin** Step 4: **For** each Dataset 'DS' with Features 'F' and pre-processed data samples 'PD' Step 5: Evaluate skewness 'Sk' according to (8) Step 6: Evaluate kurtosis 'Ku' according to (9) Step 7: **If** '$Sk > 0$' Step 8: **Then** the feature distribution is right-skewed Step 9: Consider the resultant 'Sk' value for conducting Jarque–Bera Test Step 10: **End if**

Step 11: **If** '$Sk < 0$'

Step 12: **Then** the feature distribution is left-skewed

Step 13: **Discard** the resultant 'Sk' value and go to step 5

Step 14: **End if**

Step 15: Formulate Jarque–Bera Test according to (10)

Step 16: **If** '$JB \geq 0$ *and* $JB < 0.5$'

Step 17: **Then** features are selected

Step 18: **Return** features selected 'FS'

Step 19: **End if**

Step 20: **If** '$JB \geq 0.5$ *and* $JB < 1$'

Step 21: **Then** features are discarded

Step 22: **Go to** step 4

Step 23: **End if**

Step 24: **Return**

Step 25: **End for**

Step 26: **End**

Algorithm 2 Weight Jarque–Bera Kurtosis Matching Regressive Feature Selection

4.Experimental setup

Proposed, SNDW-KMR model and existing (GDMPredictor) [1] and (MRA-optimized SVM) [2] are implemented in Python. The pregnancy risk factor dataset obtained from https://www.kaggle.com/datasets/mmhossain/pregnancy-risk-factor-data.

2.4 Performance metrics of data pre-processing

Performing pre-processing (PT): Time consumed for data pre-processing. $PT = \sum_{i=1}^{m} S_i * Time\ (PD)$ (11)

In (11), data samples 'S_i' and time consumed to preprocess '$Time\ (PD)$'. It is measured in terms of seconds (sec). Error rate is defined as a amount of samples to undetected.

$$ER = \sum_{i=1}^{m} \frac{S_{IO}}{S_i} * 100 \quad (12)$$

From, Error rate 'ER' is measured and producing inaccurate outcomes 'S_{IO}'.

2.5 Performance metrics of feature selection

Two distinct feature selection performance metrics are analyzed and validated.

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (14)$$

True positive 'TP', true negative rate 'TN' false positive 'FP', false negative 'FN' It is measured in terms of percentage (%).
False positive rate (FPR**)**: the analysis of maternal health care prediction is evaluated as,

$$FPR = \frac{FP}{FP+TN} \quad (15)$$

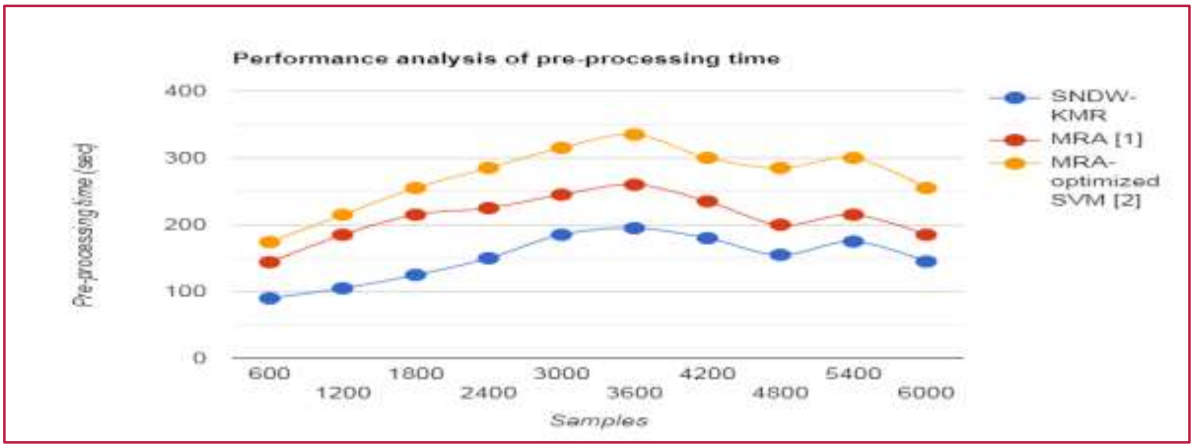

Figure 4 Graphical representation of pre-processing time

In figure 4, illustrates pre-processing time of SNDW-KMR reduced by 29% [1] and 45% [2].

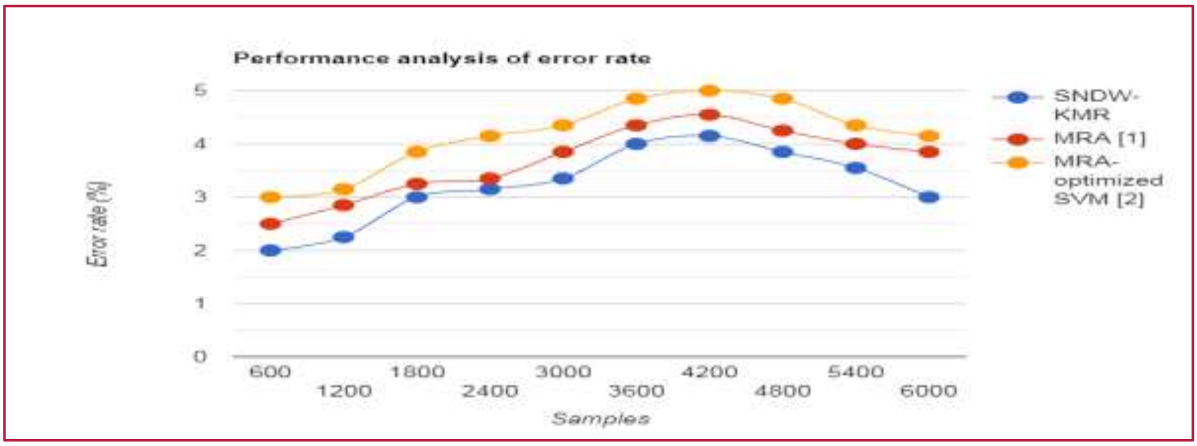

Figure 5 Graphical representation of error rate

Figure 5 illustrates error of SNDW-KMR method by reduced 12% [1] and 23% [2]

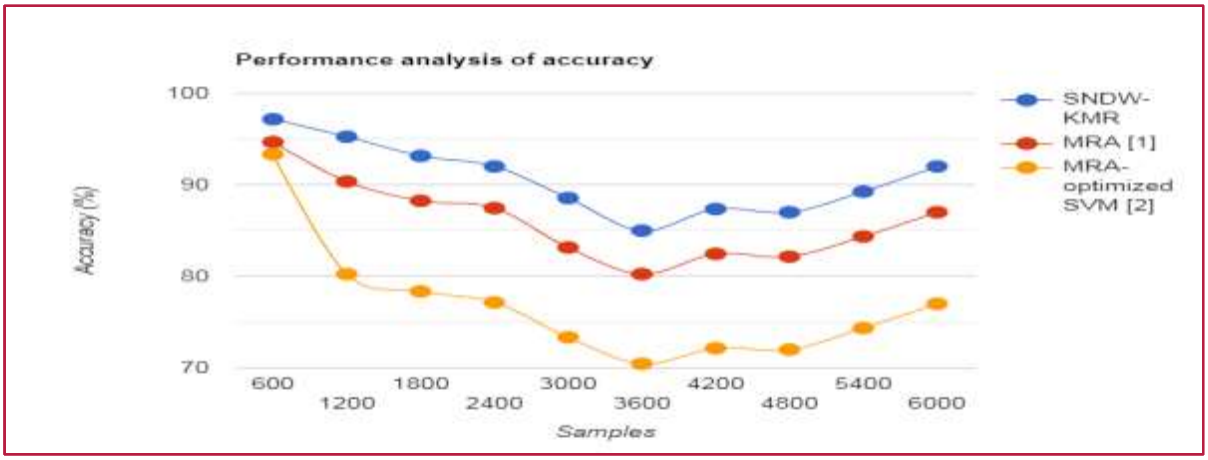

Figure 6 Graphical representation of accuracy

Figure 6 illustrates accuracy of SNDW-KMR method improved by 5% [1] and 18% [2]

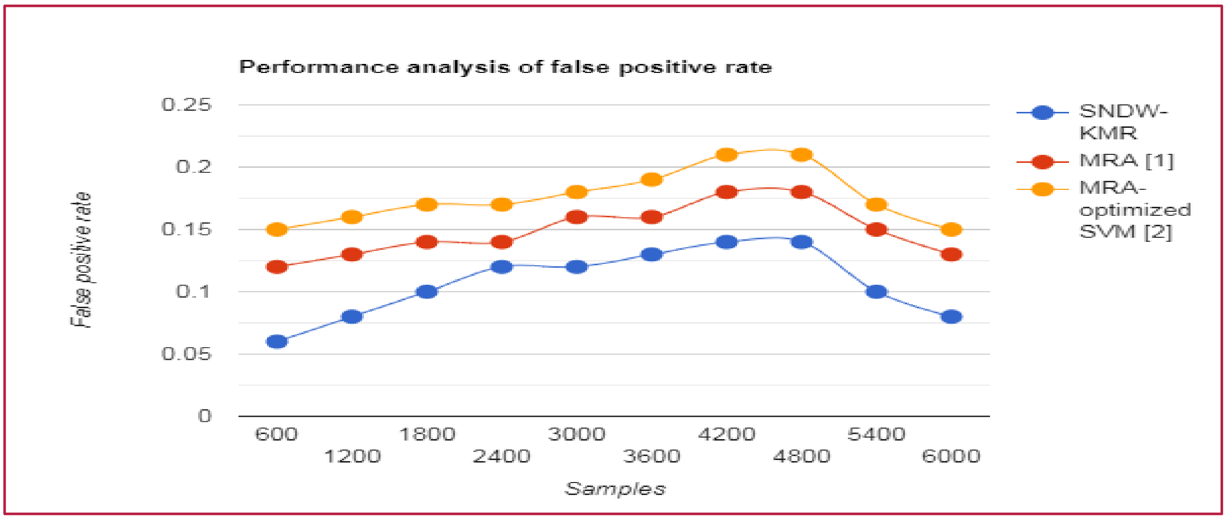

Figure 7 Graphical representation of false positive rate

Finally, figure 7 false positive rate using SNDW-KMR method by 29% [1] and 40% [2]

3. CONCLUSION

The SNDW-KMR for maternal health risk prediction is proposed. Proposed method realizes the joint modeling and analysis of maternal health risk prediction and improve identification performance. Simulation results using SNDW-KMR method provides attractive benefits upon comparison to existing methods with pre-processing time, error rate, accuracy and false positive rate

REFERENCES

[1] J. Xing, K. Dong, X. Liu, J. Ma, E. Yuan, L. Zhang and Y. Fang, "Enhancing gestational diabetes mellitus risk assessment and treatment through GDMPredictor: a machine learning approach", Journal of Endocrinological Investigation, Springer, March 2024, Pages 1-10 [gestational diabetes mellitus predictor (GDMPredictor)]

[2] Abeer S. Desuky, Sadiq Hussain, Mehmet Akif Cifci, Lamiaa M. El Bakrawy, Olfa Mzoughi, Naoufel Kraiem, "Parameter Optimization Based Mud Ring Algorithm for Improving the Maternal Health Risk Prediction", IEEE Access, Vol. 12, Nov 2024

[3] Claudio Michael Louis SKom, Nining Handayani, MBiomed, Tri Aprilliana, Arie A. Polim, Arief Boediono and Ivan Sini, "Genetic algorithm–assisted machine learning for clinical pregnancy prediction in vitro fertilization", AJOG Global Reports, Elsevier, Volume 3, Issue 1, February 2023, Pages 1-5

[4] Chetanya Puri, Gerben Kooijman, Felipe Masculo, Shannon Van Sambeek, Sebastiaan Den Boer, Jo Hua, Nan Huang Henry Ma Yafang Jin, Fan Ling Guanghui Li, Dongtao Zhang Xiaochun Wang, Stijn Luca, Bart Vanrumste, "A Personalized Bayesian Approach for Early Intervention in Gestational Weight Gain Management toward Pregnancy Care", IEEE Access, Volume 9, November 2021, Pages 160946 – 160957

[5] Md Shahin Ali, Md Maruf Hossain, Moutushi Akter Kona, Kazi Rubaya Nowrin and Md Khairul Islam, "An ensemble classification approach for cervical cancer prediction using behavioral risk factors", Healthcare Analytics, Elsevier, Volume 5, June 2024, Pages 1-15

[6] Yuanxin Yao, Rongjie Liu and Bo Zhang, "FetalAI: A deep learning web-based application for predicting birthweight from prenatal ultrasound measurements", Informatics in Medicine Unlocked, Elsevier, Volume 49, 2024, Pages 1-12

[7] Salman Zahid, Shikha Jha, Gurleen Kaur, Youn-Hoa Jung, Anum S. Minhas, MHS, Allison G. Hays, Erin D. Michos, "A Machine Learning Risk-Prediction Model for Acute Peripartum Cardiovascular Complications During Delivery Admissions", CARDIO-OBSTETRICS, National Institute of Health, Elsevier, Aug 2024

[8] Chen Wang, Anna L. V. Johansson, Cina Nyberg, Anuj Pareek, Catarina Almqvist, Sonia Hernandez-Diaz, Anna S. Oberg, "Prediction of pregnancy-related complications in women undergoing assisted reproduction, using machine learning methods", National Institute of Health, Elsevier, Feb 2024

[9] Xiaoshi Zhou, Feifei Cai, Shiran Li, Guolin Li, Changji Zhang, Jingxian Xie, Yong Yang, "Machine learning techniques for prediction in pregnancy complicated by autoimmune rheumatic diseases: Applications and

challenges”, International Immunopharmacology, Elsevier, Vol. 134, Jun 2024

[10] Jonathan S. Schor, Adesh Kadambi, Isabel Fulcher, Kartik K. Venkatesh, Mark A. Clapp, Senan Ebrahim, Ali Ebrahim, Timothy Wen, “Using machine learning to predict the risk of developing hypertensive disorders of pregnancy using a contemporary nulliparous cohort”, AJOC Global Reports, Elsevier, Aug 2024

[11] Mohammad Mobarak Hossain, Mohammod Abdul Kashem, Nasim Mahmud Nayan, Mohammad Asaduzzaman Chowdhury, “A Medical Cyber-physical system for predicting maternal health in developing countries using machine learning”, Healthcare Analytics, Elsevier, Nov 2023

[12] Dipti Dash, Mukesh Kumar, “An ensemble-based stage-prediction machine learning approach for classifying fetal disease”, Healthcare Analytics, Elsevier, Vol. 5, Jun 2024

[13] Yuhan Du, Catherine McNestry, Lan Wei, Anna Markella Antoniadi, Fionnuala M. McAuliffe, Catherine Mooney, “Machine learning-based clinical decision support systems for pregnancy care: A systematic review”, International Journal of Medical Informatics, Elsevier, May 2023

[14] Rosita Guido, Stefania Ferrisi, Danilo Lofaro and Domenico Conforti, “An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review”, Information, MDPI, Oct 2024

[15] Aymin Javed, Nadeem Javaid, Muhammad Hasnain Umair Sarfraz, Imran Ahmed, Muhammad Shafiq, Jin-Ghoo Choi, “Applying Advanced Data Analytics on Pregnancy Complications to Predict Miscarriage With eXplainable AI”, IEEE Access, Vol. 12, Dec 2024

[16] Yue Wu, Xixuan Yu, Mengting Li, Jing Zhu, Jun Yue, Yan Wang, Yicun Man, Chao Zhou, Rongsheng Tong, Xingwei Wu, “Risk prediction model based on machine learning for predicting miscarriage among pregnant patients with immune abnormalities”, Frontiers in Pharmacology, Vol. 15, Apr 2024

[17] Muhammad Irfan, Setio Basuki, Yufis Azhar, “Giving more insight for automatic risk prediction during pregnancy with interpretable machine learning”, Bulletin of Electrical Engineering and Informatics, Vol. 10, Jun 2021

[18] Aman Sharma, Janvi Malhotra, Shreya Sharma, Madhav Grover, Shruti, “Predicting Maternal Health Risk Using Machine Learning Models And Comparing The Performance Of Percentage Split And K-Fold Cross Validation”, International Journal of Novel Research and Development, Vol. 9, May 2024

[19] Fajar Javed, Syed Omer Gilani, Seemab Latif, Asim Waris, Mohsin Jami, Ahmed Waqas, “Predicting Risk of Antenatal Depression and Anxiety Using Multi-Layer Perceptrons and Support Vector Machines”, Journal of Personalized Medicine, MDPI, Oct 2021

[20] Tünde Montgomery-Csobán, Kimberley Kavanagh, Paul Murray, Chris Robertson, Sarah J E Barry, U Vivian Ukah, Beth A Payne, “Machine learning-enabled maternal risk assessment for women with pre-eclampsia (the PIERS-ML model): a modelling study”, Digital Health, The Lancet, Vol. 6, Apr 2024

[21] Tamar Krishnamurti, Samantha Rodriguez, Bryan Wilder, Priya Gopalan, Hyagriv N. Simhan, “Predicting first time depression onset in pregnancy: applying machine learning methods to patient-reported data”, Archives of Women's Mental Health, May 2022

[22] Sulaiman Salim Al Mashrafi, Laleh Tafakori, Mali Abdollahian, “Predicting maternal risk level using machine learning models”, BMC Pregnancy and Childbirth, Oct 2024

[23] Hiba Asri, Zahi Jarir, “Toward a smart health: big data analytics and IoT for real‑time miscarriage prediction”, Journal of Big Data, Springer, Vol. 10, Nov 2023

[24] Idris Zubairu Sadiq, Fatima Sadiq Abubakar, Muhammad Auwal Saliu, Babangida Sanusi katsayal, Aliyu Salihu, Aliyu Muhammad, “Machine learning algorithms for predictive modeling of dyslipidemia‑associated cardiovascular disease risk in pregnancy: a comparison of boosting, random forest, and decision tree regression”, Bulletin of the National Research Centre, Springer, Jan 2025

[25] Michael Muia Munyao, Elizaphan Muuro Maina, Shadrack Maina Mambo, Anthony Wanyoro, “Real‑time pre‑eclampsia prediction model based on IoT and machine learning”, Discover Internet of Things, Springer, Vol. 4, Aug 2024

[26] Issac Neha Margret, K. Rajakumar, K. V. Arulalan, S. Manikandan, Valentina, “Statistical Insights Into Machine Learning-Based Box Models for Pregnancy Care and Maternal Mortality Reduction: A Literature Survey”, IEEE Access, Vol. 12, May 2024

[27] Leila Jamel, Muhammad Umer, Oumaima Saidani, Bayan Alabduallah, Shtwai Alsubai, Farruh Ishmanov, Tai-hoon Kim, Imran Ashraf, “Improving prediction of maternal health risks using PCA features and TreeNet model”, Peer Journal of Computer Science, Apr 2024

[28] Taofeeq Oluwatosin Togunwa, Abdulhammed Opeyemi Babatunde, Khalil-ur-Rahman Abdullah, “Deep hybrid model for maternal health risk classification in pregnancy: synergy of ANN and random forest”, Frontiers in Artificial Intelligence, Vol. 6, Jul 2023

[29] Subhash Mondal, Amitava Nag, Anup Kumar Barman and Mithun Karmakar, “Machine learning-based maternal health risk prediction model for IoMT framework”, International Journal of Experimental Research and Review, Vol. 32, Aug 2023

[30] Xi Hang Cao, Ivan Stojkovic, Zoran Obradovic, “A robust data scaling algorithm to improve classification accuracies in biomedical data”, BMC Bioinformatics, Oct 2016.

..