IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 26, Issue 3, Ser. 1 (May. – June. 2024), PP 07-17
www.iosrjournals.org

Classification Of Health Risk Levels For Pregnant Women Using Support Vector Machine (SVM) Algorithm

Md. Ashikur Rahman[1], Rifat Mohammad Noor[2] , Soumya Mallik[3,]
Nurjahan Kamal Santa[4], Sayanti Deb[5], Abhijit Pathak[6]

[1]*(Department Of Computer Science & Engineering, BGC Trust University Bangladesh, Bangladesh)*
[2]*(Department Of Computer Science & Engineering, BGC Trust University Bangladesh, Bangladesh)*
[3]*(Department Of Computer Science & Engineering, BGC Trust University Bangladesh, Bangladesh)*
[4]*(Department Of Computer Science & Engineering, BGC Trust University Bangladesh, Bangladesh)*
[5]*(Department Of Computer Science & Engineering, BGC Trust University Bangladesh, Bangladesh)*
[6]*(Department Of Computer Science & Engineering, BGC Trust University Bangladesh, Bangladesh)*

Abstract:
This investigation undertakes a robust quantitative approach, centring on the utilization of the Support Vector Machine (SVM) classification algorithm to accurately predict the health risk levels of pregnant women. By employing SVM, renowned for its efficacy in classification tasks, the study aims to provide a reliable means of forecasting disease risks among expectant mothers. The initial dataset exhibited a modest accuracy rate of 60%, which saw a significant improvement to 79% after undergoing meticulous preprocessing procedures. These preprocessing steps were pivotal in addressing data imbalances inherent in the initial dataset, thereby enhancing the accuracy by a noteworthy 19%. Such enhancement underscores the importance of preprocessing in refining data quality for predictive modelling tasks. By offering a robust predictive framework, this study holds the potential to yield valuable insights for healthcare practitioners, facilitating early detection and intervention strategies tailored to the individual health risk profiles of pregnant women. This proactive approach to healthcare management can significantly contribute to improved maternal and fetal outcomes. Drawing upon a comprehensive UCI dataset comprising diverse variables such as systolic and diastolic blood pressure, glucose levels, heart rate, and risk parameters, encompassing a substantial 1014 data entries, the research ensures a comprehensive analysis of pertinent factors influencing maternal health. Through rigorous analysis and model refinement, this study endeavours to advance the capabilities of predictive analytics in maternal healthcare, ultimately enhancing the quality of care provided to pregnant women worldwide.

Keywords*: Support Vector Machine (SVM), Pregnant Women Health Risk, Predictive Modeling, Classification algorithm, Risk level prediction.*

Date of Submission: 24-04-2024 Date of Acceptance: 04-05-2024

I. Introduction

Maternal health refers to a mother's physical and mental well-being during pregnancy, birth, and postpartum. However, throughout these pregnancies, women face varied degrees of pregnancy complications, which can be damaging to both the mother's and the fetus's health. Before becoming pregnant, these problems can affect even healthy women. Maternal mortality is the term used by the World Health Organization (WHO) to describe mother deaths brought on by pregnancy-related issues. SDG3 of the UN Sustainable Development Agenda aims to eradicate maternal and newborn mortality issues. Every day, over 6,700 children and 810 pregnant women die (WHO, 2019, 2020) [11].

Major causes of these maternal deaths include difficulties such as hypertension, diabetes, preeclampsia, venous blood clotting disorders, twin and multiple pregnancies, rheumatic heart disease, bleeding, and premature birth. Low-income nations struggle with maternal mortality due to prevailing attitudes and habits, limited health care, and lack of education. Researchers have split the predicted maternal mortality into distinct death rates, such as Crude Death Rates (CDRs), age-specific death rates, and age-standardized death rates. Still, these characteristics are more common in high-income and low-income countries. But in addition to serious pregnancy complications, other risks include medical misdiagnosis and racial/ethnic health disparities [12].

Therefore, early detection of pregnancy complications before premature birth or death and regular monitoring of systolic and diastolic blood pressure, age, pulse rate, blood oxygen velocity, body temperature, and respiratory rate during pregnancy and timely awareness can reduce maternal mortality to a large extent. Machine learning methods are essential for analyzing a mother's health data and factors of complications, as well as testing and predicting risk reduction. Using models based on machine learning is believed to reduce maternal mortality through the timely detection of complications effectively. Advanced predictive analytics can revolutionize healthcare by identifying risks, reducing risk, and reducing maternal and child mortality[13].

The stage of reproduction that a woman experiences is pregnancy. Menstruation begins at puberty and signifies a woman's potential for conception. Pregnant women necessitate specialized care due to substantial physical and emotional changes. Prioritizing maternal and fetal health during childbirth mitigates complications. However, a lot of expectant mothers are still unaware of how crucial prenatal care is. Limited access to prenatal care and low health literacy reduces maternal well-being in developing nations. The mother's well-being directly impacts the health of the newborn. Hormonal changes and physical changes accompany pregnancy, which emphasizes the necessity for a specialized diet and close monitoring. Pregnant women frequently deal with diabetes, high blood pressure, anemia, depression, and mental health conditions. Furthermore, habits like smoking and alcohol consumption heighten health risks for both mother and child. Previous studies indicate that only 32.1% of respondents could self-detect disease risks [14].

The study highlights a critical problem in maternal health, specifically the frequency of pregnancy difficulties that hurt the health of the mother and fetus and ultimately increase the risk of maternal death. The issue is further worsened by differences in healthcare practices, understanding, and access, especially in low-income nations [15].

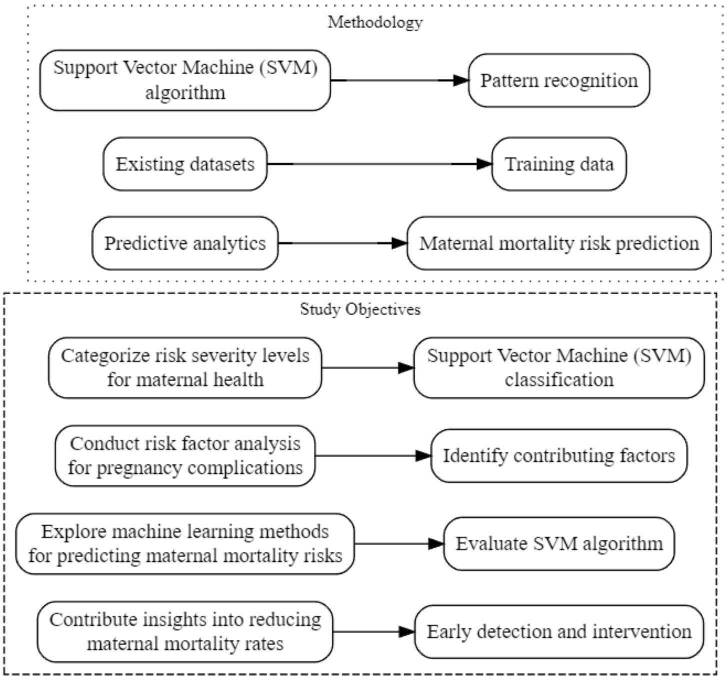

Methodology
Support Vector Machine (SVM) algorithm
Pattern recognition
Existing datasets
Training data
Predictive analytics
Maternal mortality risk prediction
Study Objectives
Categorize risk severity levels for maternal health
Support Vector Machine (SVM) classification
Conduct risk factor analysis for pregnancy complications
Identify contributing factors
Explore machine learning methods for predicting maternal mortality risks
Evaluate SVM algorithm
Contribute insights into reducing maternal mortality rates
Early detection and intervention

Figure 1: A Directed Graph Illustrating the Study's Objectives and Methodology

This study's central research question is: How can risk factor analysis and machine learning techniques, notably the Support Vector Machine (SVM) algorithm, be used to classify the degree of maternal health risks and minimize pregnancy complications and maternal death rates?

To categorize risk severity levels for maternal health by employing the Support Vector Machine (SVM) technique.

- To conduct risk factor analysis for pregnancy complications.
- To explore the potential of machine learning methods in predicting maternal mortality risks based on specific parameters.
- To contribute insights into reducing maternal mortality rates through early detection and intervention using advanced predictive analytics and machine learning techniques.

In Figure 1, a directed graph illustrates the study's objectives and methodology, with each objective linked to its corresponding methods.

The use of SVM can study the pattern of data of the patient whose diagnosis is known and can be used to predict the diagnosis of other patients based on previously learned patterns. Using existing datasets, the results of this study will be able to predict whether pregnant mothers have a high, moderate, or low risk of mortality based on specific parameters.

II. Related Works

Wanriko et al. (2021, March) explored the Risk Assessment of Pregnancy-induced Hypertension. They investigate seven machine learning algorithms based on a public dataset of Logan (2020): Logistic Regression, K-nearest Neighbor, Decision Tree, Random Forest (RF), Multi-layer Perceptron Neural Network, Support Vector Machines, and Naive Bayes. The authors performed the SMOTE algorithm data transformation, which consisted of MinMaxScaler, StandardScaler, Normalizer, and PCA on the dataset.The investigation found that RF had the highest accuracy at 89.62 per cent. There were 352 sample sizes [1]. Hossain, M. I. et al. (2022) classified Unintended Pregnancy among Married Women in Bangladesh using six machine learning algorithms, including Logistic Regression, Random Forest, and Support vector machine. K-nearest neighbours, Na¨ıve Bayes, elastic net regression combines ridge penalty and Least Absolute Shrinkage and Selection Operator. They build the predictive model using DHS cross-sectional data. The authors found that ENR produces more accurate predictions (approximately 75%) than others. They used 20,127 samples [2]. Chelsea, M. Y., & Rosa, P. H. P. (2024) classified the delivery type of pregnant women by using three kinds of SVM kernels, namely Linear, RBF, and Polynomial kernels, and classified the dataset using several variations of parameters of C, gamma, and degree.In the investigation, the authors found that the highest accuracy is 92.98% at 5-fold cross-validation using the RBF kernel with parameters C = 10 and gamma = 1. Also, they found that the performances of the three SVM kernels varied depending on the type of data used. The number of datasets was 302 [3]. Javed et al. (2021) predicted and classified factors affecting preterm delivery by comparing two statistical methods, SVM and logistic regression. To model the SMO algorithm, They used a linear Kernel. In the end, the authors found that the SVM model performs better in predicting the factors affecting premature delivery than the logistic regression model, with an accuracy rate of 66%. They used 600 datasets [4]. Mou et al. (2021) investigated the prevalence of preeclampsia and the associated risk factors among pregnant women in Bangladesh.The authors performed a Logistic regression analysis to identify preeclampsia-related factors.They also performed colourimetric, kinetic, and dipstick methods to analyze the dataset and IBM SPSS, version 25.0, for statistical data analysis. In the investigation, they found that the prevalence of preeclampsia was comparatively higher in rural areas of Bangladesh. They investigated 110 datasets [5]. Odunayo, Moududur, Anjan, Rubhana & Faruk(2022) identified problems in a prestigious sector of medical science about pregnancy, which was the Prevalence and risk factors of vitamin B12 deficiency among pregnant women in rural Bangladesh who were in their early pregnancy using the methodology based on questionnaire method and the chi-squared test. The Authors found that the study unequivocally reveals that the prevalence of vitamin B12 deficiency among these women increased significantly as their pregnancies progressed. During early pregnancy, 19% of the women had vitamin B12 deficiency, which doubled to 38% during late pregnancy. In this research, the sample size was 522 [6]. Moreira, Rodrigues, Marcondes, Neto, Kumar & Diez(2018) researched a preterm birth risk in hypertensive pregnant women prediction system for mobile health applications system, which is based on the Support Vector Machine Algorithm using the linear kernel, ANN, and quadratic optimization for identifying risk factors, symptoms, and diseases, predicting gestational age at delivery, and assessing the fetus' condition post-childbirth. The Authors propose a system architecture integrating mobile health and DSSs for efficient healthcare delivery. Using 205 samples, the SVM model achieved good accuracy (0.821) in predicting preterm childbirth risk and newborn outcomes[7]. MUTLU, DURMAZ, CENGİL, and YILDIRIM(2023) investigated the factors using machine learning to assess risks in pregnancy. The authors use six machine-learning models: decision tree algorithms, LightGBM, CatBoost, Random, Forest, GBM, and KNN [27]. The Decision Tree classifier yielded the greatest accuracy score of 89.16%, according to the authors' prediction based on 203 samples. The Light GBM classifier followed this accuracy rate at 84.24%, CatBoost at 83.74%, Random Forest at 81.28%, Gradient Boosting Machines at 73.89% and KNN at 68.47% [8]. Nurul & Sultana(2024) analyzed various kinds of health risk issues during pregnancy to prevent pregnancy-related issues by reducing the number of errors. The suggested model outperforms all others in terms of accuracy and efficiency using LDA, QDA, KNN, Decision Tree, Random Forest, Bagging, and Support Vector Machine, Cost, Gamma, with an accuracy score of 86.13% for the support vector machine using a 10-fold cross-validation technique with 800 observations as training data and 214 as test data [26]. The models utilized six predictors: age, systolic BP, diastolic BP, blood sugar, body temperature, and heart rate[9]. Taofeeq, Abdulhammed & Khalil-ur-Rahman(2023) researched a deep hybrid model for maternal health risk classification in pregnancy using ANN & RF, which is based on training(75%) and training sets(25%) with key features of health risk during pregnancy such as age, systolic and diastolic blood pressure, blood sugar, body temperature, and heart rate. The results of this model achieved 95% accuracy, 97% precision, 97% recall, and an F1 score of 0.97 on the testing dataset[10].

III. Methodology

This study uses quantitative methods. This method is more systematic and should use data, numbers, or measurable variables. Quantitative research aims to develop structured, mathematical, and theoretical models that relate to the problems to be thoroughly examined. Thus, it can be concluded that the quantitation method refers to data, theory, and hypothesis. This method is appropriate for this study because it uses datasets that contain data on diseases that threaten the health factors of pregnant mothers. This study has several stages. The stages can be seen in Figure 2.

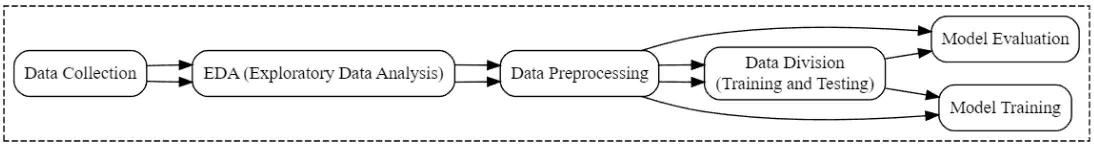


Figure 2: A Directed Graph Illustrating the Sequential Phases of the Study, from Data Collection through Model Evaluation

In Figure 2, the first phase is the collection of datasets. The data sets obtained will then enter the EDA (Exploration Data Analysis) phase, which aims to understand the contents of the data sets. After passing through EDA, then go to the preprocessing phase of data. At the data preprocessing stage, the data will be processed through several phases and enter the data division into data training and data testing. When the data is shared, then the data goes into the model training and evaluation phase [16].

Data Collection

This study's data set comes from the UCI data set. The data includes some parameters that affect the health of pregnant mothers. Some of the data can be seen in Table 1.

Table 1. Tabel Dataset

Sl. No	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
0	25	130	80	15.0	98.0	86	High Risk
1	35	140	90	13.0	98.0	70	High Risk
2	29	90	70	8.0	100.0	80	High Risk
3	30	140	85	7.0	98.0	70	High Risk
4	35	120	60	6.1	98.0	76	Low Risk
...	...	...	...	...	...	...	...
1009	22	120	60	15.0	98.0	80	High Risk
1010	55	120	90	18.0	98.0	60	High Risk
1011	35	85	60	19.0	98.0	86	High Risk
1012	43	120	90	18.0	98.0	70	High Risk
1013	32	120	65	6.0	101.0	76	Mid Risk

Table 1 is a sample dataset containing data from five people with several health parameters such as age, systolicBP, diastolicBP, BS, body temperature, heart rate, and risk level. The dataset has seven columns and 1014 lines. The age column is the column of the patient's age, the SystolicBP column and the DiastolicBPs column are the columns of the systolic blood pressure and dystolic blood pressure in mmHg units, the BS column represents the blood sugar level of a patient in mmol/L, the BodyTemp column presents the body temperature in degrees Fahrenheit, the HeartRate column the heart rate in beats per minute in a patient, and the RiskLevel column, the risk level column for the health of a pregnant woman. Research can be conducted utilizing datasets and various parameters to classify the health status of pregnant mothers, aiming to ascertain their level of health risk during pregnancy [17].

EDA (Exploration Data Analysis)

EDA is the process of exploring and understanding datasets, starting from the relationship between columns, amounts of data, and data distribution. It aims to understand the data sets better so they can be processed accurately. At this stage of EDA, the contents in data sets, such as outliers, data duplication, missing value, encoding, and noisy data, can be known. By knowing that data and processing it, the data and results will be cleaner [18]. There are several techniques in the application of the EDA, among others:

- *Data collection:* The data will be processed in the input at this stage.
- *Viewing the data structure:* Viewing data structure aims to understand the data to be thoroughly examined.
- *Data selection:* The purpose of this data selection is to remove data that is not important and has no influence on the modelling process.
- *Changing variables:* This purpose works to ensure that data modelling usually runs.
- *Find outlier and null data:* This aims to see vulnerable, abnormal, and empty data.

Preprocessing Data

Preprocessing data is an essential step in processing data in a study. This preprocessing aims to clear the data, process it, and prepare it for subsequent processing. By carrying out the preprocessing phase, raw data will become more formal and accessible to process, thus making it more effective for further processing. The first phase is data cleaning. This is a phase of data cleansing aimed at removing data that is at risk of reducing the level of accuracy. Once the data cleaning phase is complete, integration is the next stage [19]. At this stage, the data obtained is combined from various sources and then merged into one. The next stage is the stage of data transformation. At this phase, the data will be harmonized; the purpose of harmonizing this data is so that the data can be appropriately processed. The last step is data reduction. At this stage, the data will be reduced. This phase aims to reduce the amount of data samples taken. In this preprocessing process, the data used in this study had no empty or null data. The data labelled in the Risklevel column was modified with numerical parameters so that the data could be processed. A table of data sets that have already been preprocessed can be seen in Table 2.

Table 2. After Preprocessing

Sl. No	Risk Level	Count
0	Low Risk	406
1	Mid Risk	336
2	High Risk	272

SVM (Support Vector Machine)

SVM is one of the algorithms that is often used to analyze data and sort data. This SVM algorithm is frequently used in research, especially in disease classification. The SVM is divided into two, namely, a linear SVM and a non-linear SVM. This algorithm belongs to the supervised learning category. Supervised Learning is when the obtained data already has a label and remains to be processed. On the SVM algorithm, there is a trick kernel. Kernel trick is a method to change data in a particular dimension, e.g., change the 2D dimension to 3D. The change of the dimension is aimed at making the hyper line more optimally produced. There are several kernel functions: linear kernel, RBF, polynomial, and sigmoid [20].

Key Concepts and Potential for Pregnancy Risk Prediction

- **Linear Kernel Function (K(x, y) = x · y):** This formula defines the similarity between two data points (x and y) in the SVM. The dot product (x · y) captures how closely aligned the data points are in a high-dimensional space. In pregnancy risk prediction, similar data points might represent mothers with comparable risk profiles based on their features (e.g., blood pressure readings).
- **Linear SVM Decision Function (f(x) = sgn(w.x + b)):** This equation defines the decision function, which calculates a linear separation between the high-risk and low-risk classes based on the features (x) of a pregnant mother. Here:
  - **w**: A weight vector learned by the SVM during training. It reflects the importance of each feature in determining risk.
  - **x**: A vector representing a mother's features (e.g., blood pressure, age).
  - **b**: The bias term that helps position the decision boundary.
  - **sgn( )**: Signum function, which outputs +1 if the expression is positive (indicating *high-risk) and -1 if negative (indicating low-risk).*

How SVM can contribute to optimal risk prediction

- **Classification:** By applying the decision function (f(x)) to a mother's features, the SVM can classify her into a risk category (high, medium, low) based on the output of the function. This classification facilitates early intervention for pregnancies at high risk.
- **Feature Importance:** Analyzing the weight vector (w) can reveal which features (e.g., blood sugar) are most influential in the decision function, helping healthcare providers prioritize monitoring specific factors.
- **Clear Decision Boundary:** Although representing a linear separation, the decision function can be visualized to understand how different features contribute to risk classification. This can be valuable for interpreting the model and potentially improving its generalizability [21].

Confusion Matrix

A Confusion Matrix is a table with several combinations of the value of the classification result. The confusion matrix has the following four terms: False Positive (FN), False Negative (TN), True Positive (TP), and False Negative (FN). Data that is damaging but is mistakenly identified as positive is called a false positive (FP) [22]. An overview of the confusion matrix table can be found in Table 3.

Tabel 3. Confusion Matrix

		True	**False**
Predicted Value	**True**	TP (True Positive)	FP (False Positive)
	False	FN (False Negative)	TN (True Negative)

From the table 3, accuracy, precision, recall, and F-1 score values can be obtained. Accuracy values are values that determine how accurate the result of the classification is; the values can be calculated with the formula [25]:

$$\textbf{Accuracy} = \left(\frac{(TP*TN)}{(\boldsymbol{TP*FP*TN*FN})}\right) \textbf{ x 100\%}$$

Precision is the accuracy value between your data and the prediction result.

$$\textbf{Precision} = \left(\frac{TP}{(\boldsymbol{TP*FP})}\right) \textbf{ x 100\%}$$

Recall is a value that describes the model's success in finding information. The recall value can be calculated using the formula:

$$\textbf{Recall} = \left(\frac{TP}{(\boldsymbol{TP*FN})}\right) \textbf{ x 100\%}$$

The F-1 score is a value that compares the average of the precision value and the recall value. This value can be calculated using the formula:

$$\textbf{F-1 score} = \left(\frac{\boldsymbol{2\ x\ precision\ x\ recall}}{(\boldsymbol{precision*recall})}\right) \textbf{ x 100\%}$$

IV. Results And Discussion

After going through the EDA, preprocessing, and modelling stages, the next stage is model testing. The testing of this model was carried out to determine the performance of the model in terms of the classification of the health risk of the pregnant mother that was produced. The accuracy of the results obtained in this data modelling is 79%. A model with a 79% accuracy means that the data modelling can correctly predict as much as 79% of all the data being evaluated.

In addition to accuracy, other matrices are used to view modelling results. In the confusion matrix table, several parameters are used in this study: high risk, medium risk, and low risk. The confusion matrices help look at this study's more complex accuracy picture; the resulting confusion matrix can be seen in Table 4.

Table 4. Confusion Matrix Testing

	High risk	**Medium risk**	**Low risk**
High risk	75	10	2
Medium risk	39	37	5
Low risk	4	10	62

The correct sample indicates a high risk of 75 on the high-risk variable. However, ten samples were wrongly predicted as medium risk, and two samples were predicted as low risk. Furthermore, the model can expect that the correct sample shows a medium risk of 37 on the medium risk variable. However, some models are incorrectly predicted as high risk by 39 and wrongly foresee low risk by 5. By knowing some parameters and results obtained from the confusion matrix tables, the results obtained and obtained can be scanned, the accuracy comparison can be learned, and the number of samples predicted correctly can be seen.

After obtaining the results of the confusion matrix table, the results are then summarized or reported using the classification report. This classification report will show some parameters and results. Parameters in this classification report include precision, recall, f1-score, support, and accuracy. The confusion matrix is shown in Figure 3.

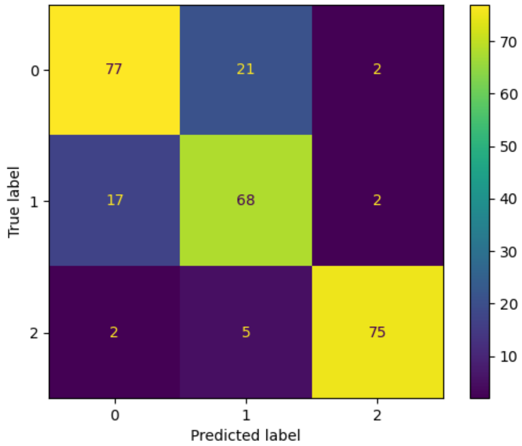


Figure 3: Confusuion Matrix

By looking at the result of the classification report in Table 5, each variable will be shown its accuracy in each parameter.

Table 5. Classification Report

	Precision	Recall	F1-score	Support
Low Risk	0.78	0.76	0.77	100
Medium Risk	0.70	0.75	0.72	87
High Risk	0.94	0.90	0.92	82
Accuracy				
Macro avg	0.81	0.80	0.80	269
Weighted avg	0.80	0.80	0.80	269

The categorization report analyzes the model's performance in detail for each of the three risk categories: "Low Risk," "Medium Risk," and "High Risk." The model's predictive capacity at varying risk levels is disclosed through detailed explanations of each category's precision, recall, and F1 score. The authors start by examining accuracy, which gauges the success rate of optimistic forecasts. The model does remarkably well in each category. It is noteworthy that it receives precision scores of 0.78, 0.70, and 0.94 for the Low, Medium, and High-Risk categories, respectively. These numbers show how well the model can recognize examples that fall into each risk category. Next, they look at recall, an indicator of the model's capacity to identify good examples, and find a similarly impressive result. Recall scores for the Low, Medium, and High-Risk categories are 0.76, 0.75, and 0.90, respectively, indicating the model's resilience in correctly identifying cases with different risk levels. F1 scores, which balance recall and precision, highlight the model's efficacy even more. For Low, Medium, and High Risk, the F1 scores are 0.77, 0.72, and 0.92, respectively, indicating how effectively the model balances reaching recall and precision targets across all categories [24]. The support column gives the model's performance metrics context by showing the number of occurrences for each risk category. The table also shows accuracy levels; weighted averages and macro averages indicate the model's overall performance. The weighted average, which considers class imbalances, produces an accuracy of 0.80, somewhat lower but still acceptable than the macro average, which considers each category equally and delivers an accuracy of 0.81.

Table 6. Model Performance Metrics Comparison

	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision	Train F1	Test F1
Low Risk	90.89456	79.92565	90.98712	76.00000	87.60330	78.35051	89.26315	77.15736
Medium Risk	90.89456	79.92565	85.71428	74.71264	87.00000	69.89247	86.35235	72.22222
High Risk	90.89456	79.92565	96.31578	90.24392	99.45652	93.67088	97.86096	91.92546

Table 6 compares performance indicators of a machine learning model during its training and testing phases, with an emphasis on three distinct risk categories: "Low Risk," "Medium Risk," and "High Risk." Each indicator provides information on how well the model generalizes from test data that hasn't been seen before to training data.

Test Accuracy: In all risk categories, the model's test accuracies fall between roughly 79.93% and 79.93%. Out of all the examples in the test dataset, this shows the percentage of correctly identified instances. The model performs consistently with low change in test accuracy across different risk categories.

Train Accuracy: Likewise, train accuracies, which range from roughly 90.89% to 90.89%, are constant throughout all risk categories. Train accuracy is the percentage of cases in the training dataset that are correctly classified out of all the instances in the dataset. The model's capacity to efficiently learn from and fit the training data is demonstrated by its excellent accuracy during the training phase.

Test Recall: This metric assesses the model's capacity to distinguish all of the positive examples in the test dataset correctly. The recall values that the model attains in various risk categories range from roughly 76.00% to 90.24%. These values suggest the model's effectiveness in capturing positive instances, with higher recall values indicating better performance.

Train Recall: Similarly, train recall values are high across all risk categories, ranging from approximately 85.71% to 96.32%. This indicates that the model effectively captures positive instances during the training phase, demonstrating its ability to learn from the training data.

Test Precision: Test precision quantifies the accuracy of optimistic predictions made by the model on the test dataset. The model achieves precision values ranging from approximately 69.89% to 93.67% across different risk categories. Higher precision values indicate fewer false optimistic predictions made by the model.

Train Precision: Train precision values remain consistently high across all risk categories, ranging from approximately 87.60% to 99.46%. This suggests that the model makes accurate optimistic predictions during the training phase, with minimal false positive predictions.

Test F1-score: A measure of the model's performance on the test dataset, the F1-score is the harmonic mean of precision and recall. The algorithm produces F1 scores in several risk categories that range from roughly 72.22% to 91.93%.

Train F1-score: In a similar vein, train F1-scores, which range from roughly 86.35% to 97.86%, continue to be high across all risk categories. These values indicate a balance between precision and recall during the training phase, with higher F1 scores indicating better overall performance.

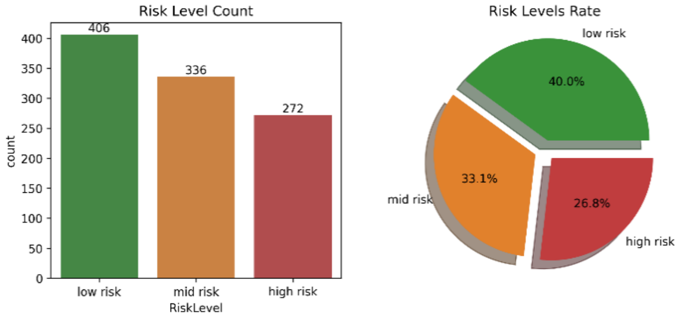
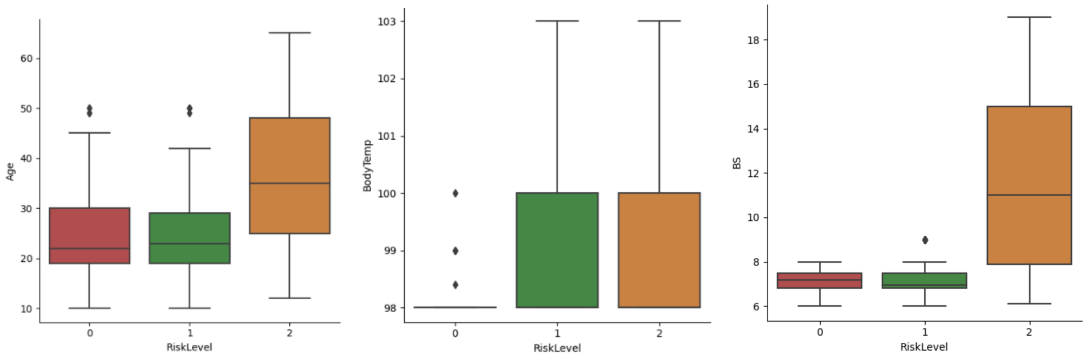


Figure 4: Distribution of Risk Levels: Frequency of Low, Medium, and High-Risk Instances in the Dataset

An illustration of risk levels is shown in Figure 4, which sheds light on how various risk categories are distributed throughout the dataset. Plotting shows how often or what percentage of cases fall into each risk category, which are "Low Risk," "Medium Risk," and "High Risk." The various danger levels are labelled on the x-axis for simple identification and comparison. The frequency or percentage of occurrences corresponding to each risk category is shown on the y-axis. Depending on the visualization approach selected, the plot can look like a bar chart, histogram, or pie chart. Whatever its exact shape, the plot provides essential insights into the frequency of various risk categories by graphically communicating the distribution of risk levels within the dataset [23].



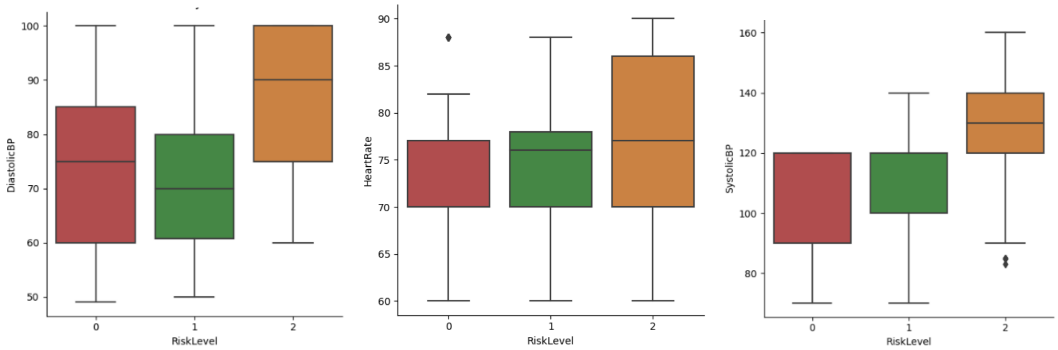


Figure 5: Visualization of Risk Level Distribution: Frequency of Low, Medium, and High-Risk Instances

Figure 5 is a visual representation of the distribution of risk levels within our dataset, shedding light on the frequency of instances categorized as low, medium, and high risk. The x-axis conveniently labels the different risk levels, enabling effortless comparison across categories. Each bar on the plot denotes the frequency of instances corresponding to a specific risk level, offering a comprehensive view of the dataset's risk distribution. This graphic is an essential resource for comprehending the frequency of various risk categories in our dataset. The frequency of Low, Medium, and High-risk occurrences can provide important information that can guide risk assessment techniques and decision-making procedures. To create efficient risk mitigation strategies and guarantee well-informed business decisions, it is essential to comprehend the distribution of risk levels.

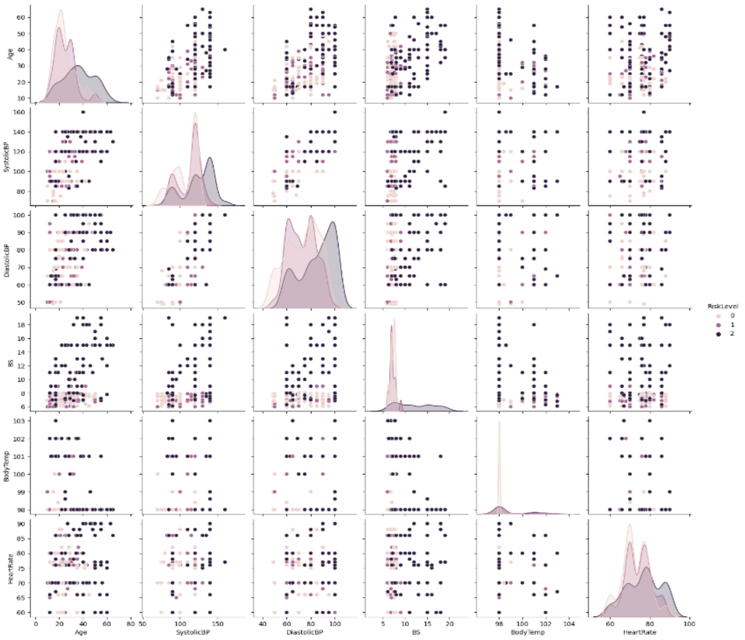


Figure 6: Pair Plot Analysis of Risk Levels: Exploring Feature Relationships Across Low, Medium, and High-Risk Categories

In Figure 6, a pair plot representing different risk levels is presented. This pair plot provides a visual overview of the relationships between various features within each risk category ("Low Risk," "Medium Risk," and "High Risk"). Each subplot in the pair plot matrix represents a combination of two features, with the diagonal subplots displaying the distribution of each feature for the respective risk category. The off-diagonal subplots display scatter plots showing the relationship between pairs of features, with each point representing a data instance coloured according to its risk level. The pair plot enables us to observe patterns, correlations, or differences in feature distributions and relationships across risk categories. By comparing the subplots for each risk level, the authors can identify potential distinctions in feature distributions or relationships that may indicate varying risk levels.

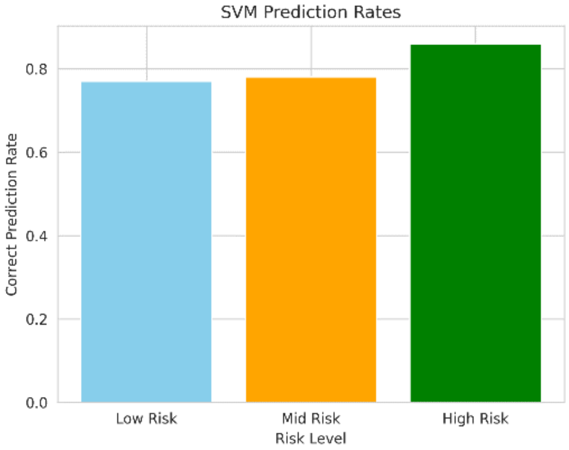


Figure 7: Risk Prediction Rates Using SVM: Comparative Analysis of Accuracy Across Low, Mid, and High-Risk Levels.

In Figure 7, a bar chart depicting risk prediction rates using SVM. The chart shows different levels of prediction accuracy, with the average and highest rates highlighted in different colours (yellow and blue). The x-axis represents the risk level (low, mid, high), while the y-axis shows the prediction rates ranging from 0 to 0.8.

V. Conclusion

The research findings indicate that employing classification methods, notably Support Vector Machine (SVM) algorithms, facilitates the prediction of health risk levels among pregnant women. Before preprocessing, the accuracy of the data stands at 60%, which notably improves to 79% post-preprocessing. This enhancement is attributed to addressing data imbalance issues inherent in the initial dataset. Consequently, the research significantly aids healthcare professionals in mitigating the mortality rate of pregnant women by enabling early detection of disease risk factors through specific parameters. Recommendations for future research include exploring alternative classification algorithms to ascertain the potential for achieving optimal model accuracy.
Moreover, leveraging multiple datasets could further enhance the accuracy and robustness of the predictive model. Despite the promising results obtained in predicting health risk levels among pregnant women using Support Vector Machine (SVM) algorithms, several limitations were identified in the research. One major constraint was the exclusive reliance on SVM algorithms for classification, potentially overlooking alternative modelling approaches that could improve predictive accuracy. Furthermore, the existence of data imbalance in the original dataset might have compromised the resilience of the prediction model despite efforts to address this problem using preprocessing approaches. In addition, the study's dependence on a restricted dataset may limit the applicability of the findings to larger populations or varied healthcare environments. Future research should look into different algorithms and make use of more enormous datasets to get over these limitations. This may lead to the development of more effective prediction models for assessing health risks associated with pregnancy.

References

[1] Wanriko, S., Hnoohom, N., Wongpatikaseree, K., Jitpattanakul, A., & Musigavong, O. (2021, March). Risk Assessment Of Pregnancy-Induced Hypertension Using A Machine Learning Approach. In 2021 Joint International Conference On Digital Arts, Media And Technology With Ecti Northern Section Conference On Electrical, Electronics, Computer, And Telecommunication Engineering (Pp. 233-237). Ieee.(Doi: 10.1109/Ectidamtncon51128.2021.9425764).

[2] Hossain, M. I., Habib, M. J., Saleheen, A. A. S., Kamruzzaman, M., Rahman, A., Roy, S., ... & Rukon, M. R. (2022). Performance Evaluation Of Machine Learning Algorithm For Classification Of Unintended Pregnancy Among Married Women In Bangladesh. Journal Of Healthcare Engineering, 2022. (Doi: 10.1155/2022/1460908).

[3] Chelsea, M. Y., & Rosa, P. H. P. (2024). Classification Of Delivery Type Of Pregnant Women Using Support Vector Machine. In E3s Web Of Conferences (Vol. 475, P. 02015). Edp Sciences. (Doi:Https://Doi.Org/10.1051/E3sconf/202447502015).

[4] Javed, F., Gilani, S. O., Latif, S., Waris, A., Jamil, M., & Waqas, A. (2021). Predicting Risk Of Antenatal Depression And Anxiety Using Multi-Layer Perceptrons And Support Vector Machines. Journal Of Personalized Medicine, 11(3), 199. (Doi: Https://Doi.Org/10.3390/Jpm11030199).

[5] Mou, A. D., Barman, Z., Hasan, M., Miah, R., Hafsa, J. M., Das Trisha, A., & Ali, N. (2021). Prevalence Of Preeclampsia And The Associated Risk Factors Among Pregnant Women In Bangladesh. Scientific Reports, 11(1), 21339.(Doi: Https://Doi.Org/10.1038/S41598-021-00839-W).

[6] Sobowale, O. I., Khan, M. R., Roy, A. K., Raqib, R., & Ahmed, F. (2022). Prevalence And Risk Factors Of Vitamin B12 Deficiency Among Pregnant Women In Rural Bangladesh. Nutrients, 14(10), 1993.(Https://Doi.Org/10.3390/Nu14101993).

[7] Moreira, M. W., Rodrigues, J. J., Marcondes, G. A., Neto, A. J. V., Kumar, N., & Diez, I. D. L. T. (2018, May). A Preterm Birth Risk Prediction System For Mobile Health Applications Based On The Support Vector Machine Algorithm. In 2018 Ieee International Conference On Communications (Icc) (Pp. 1-5). Ieee.1-5, //Doi: 10.1109/Icc.2018.8422616.

[8] Mutlu, H. B., Durmaz, F., Yücel, N., Cengil, E., & Yildirim, M. (2023). Prediction Of Maternal Health Risk With Traditional Machine Learning Methods. Naturengs, 4(1), 16-23.(Https://Doi.Org/10.46572/Naturengs.1293185).

[9] Raihen, M. N., & Akter, S. (2024). Comparative Assessment Of Several Effective Machine Learning Classification Methods For Maternal Health Risk. Computational Journal Of Mathematical And Statistical Sciences, 3(1), 161-176.(10.21608/Cjmss.2024.259490.1036).

[10] Togunwa, T. O., Babatunde, A. O., & Abdullah, K. U. R. (2023). Deep Hybrid Model For Maternal Health Risk Classification In Pregnancy: Synergy Of Ann And Random Forest. Frontiers In Artificial Intelligence, 6, 1213436.(Https://Doi.Org/10.3389/Frai.2023.1213436).

[11] Abhijit Pathak, Abrar Hossain Tasin, Ayesha Akther Esho, Ashibur Rahman Munna And Tahia Chowdhury (2020); A Smart Semi-Autonomous Fire Extinguish Quadcopter: Future Of Bangladesh Int. J. Of Adv. Res. 8 (Apr). 01-15] (Issn 2320-5407).

[12] Abiyyu, Ahmad Syafiq, And Kemas Muslim Lhaksmana. "Perbandingan Metode Seleksi Fitur Untuk Mengoptimasi Model Support Vector Machine Dalam Memprediksi Turnover Pegawai." Eproceedings Of Engineering 10.2 (2023).

[13] Afrilia, Eka, Siti Mardhatillah Musa, And Murni Lestari. "Metode Hypnosis Dalam Mengatasi Perubahan Psikologis Selama Masa Kehamilan: Studi Literatur." Jurnal Jkft 7.1 (2022): 54-58. [4] Awang Hendrianto Pratomo, W. Kaswidjanti, And S. Mu’arifah, “Implementasi Algoritma Region Of Interest (Roi) Untuk Meningkatkan Performa Algoritma Deteksi Dan Klasifikasi Kendaraan,” J. Teknol. Inf. Dan Ilmu Komput., Vol. 7, No. 1, Pp. 155– 162, 2020.

[14] Alita, Debby, Yusra Fernando, And Heni Sulistiani. "Implementasi Algoritma Multiclass Svm Pada Opini Publik Berbahasa Indonesia Di Twitter." Jurnal Tekno Kompak 14.2 (2020): 86-91.

[15] Darwis, Dedi, Eka Shintya Pratiwi, And A. Ferico Octaviansyah Pasaribu. "Penerapan Algoritma Svm Untuk Analisis Sentimen Pada Data Twitter Komisi Pemberantasan Korupsi Republik Indonesia." Jurnal Ilmiah Edutic: Pendidikan Dan Informatika 7.1 (2020): 1-11.

[16] Elfiyani, Nur Khotimah, Et Al. "Dampak Dan Strategi Layanan Kesehatan Ibu Hamil Selama Pandemi Covid-19." Jurnal Kesehatan Reproduksi 9.2.

[17] Fitriani, Diana. "Penerapan Metode Kuantitatif Dalam Penelitian Ilmiah Mahasiswa." Snpmas: Seminar Nasional Pengabdian Pada Masyarakat. 2019.

[18] Hovi, Hovi Sohibul Wafa, Asep Id Hadiana, And Fajri Rakhmat Umbara. "Prediksi Penyakit Diabetes Menggunakan Algoritma Support Vector Machine (Svm)." Informatics And Digital Expert (Index) 4.1 (2022): 40-45.

[19] Ida, Andi Syintha, And Afriani Afriani. "Pengaruh Edukasi Kelas Ibu Hamil Terhadap Kemampuan Dalam Deteksi Dini Komplikasi Kehamilan." Jurnal Inovasi Penelitian 2.2 (2021): 345-350.

[20] Juwitasari, Juwitasari, And Marni Marni. "Hubungan Antara Pengetahuan Tentang Kehamilan Resiko Tinggi Dan Tingkat Depresi Pada Ibu Hamil." Journal Of Borneo Holistic Health 3.2 (2020): 159-168.

[21] Neloy, M. A. I. ., Barua, V. ., Das, M. ., Barua, P. ., Rahat, S. U. ., & Pathak, A. . (2020). An Intelligent Obstacle And Edge Recognition System Using Bug Algorithm. American Scientific Research Journal For Engineering, Technology, And Sciences, 64(1), 133–143. Retrieved From Https://Asrjetsjournal.Org/Index.Php/American_Scientific_Journal/Article/View/5566

[22] Parapat, Indri Monika. Penerapan Metode Support Vector Machine (Svm) Pada Klasifikasi Penyimpangan Tumbuh Kembang Anak. Diss. Universitas Brawijaya, 2018.

[23] Radhi, Muhammad, Et Al. "Analisis Big Data Dengan Metode Exploratory Data Analysis (Eda) Dan Metode Visualisasi Menggunakan Jupyter Notebook." Jurnal Sistem Informasi Dan Ilmu Komputer Prima (Jusikom Prima) 4.2 (2021): 23-27.

[24] Rahmawati, Erna, Asriya Naro Rimasari, And Elvira Rm Monita. "Penyuluhan Hipertensi, Pengecekan Tekanan Darah, Kadar Gula Dalam Darah, Kolesterol Serta Asam Urat." Journal Of Community Engagement And Empowerment 1.2 (2019).

[25] Ropikoh, Isnin Apriyatin, Et Al. "Penerapan Algoritma Support Vector Machine (Svm) Untuk Klasifikasi Berita Hoax Covid-19." Journal Of Applied Informatics And Computing 5.1 (2021): 64-73.

[26] Tommy, Tommy, And Amir Mahmud Husein. "Model Prediksi Prestasi Mahasiswa Berdasarkan Evaluasi Pembelajaran Menggunakan Pendekatan Data Science." Data Sciences Indonesia (Dsi) 1.1 (2021): 14-20.

[27] Yazia, Velga, And Ulfa Suryani. "Faktor Yang Berhubungan Dengan Tingkat Stres Pada Ibu Hamil Dalam Menghadapi Persalinan." Jurnal Keperawatan Jiwa 10.4 (2022): 837-856.