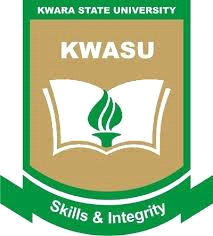
BRAIN TUMOR RECOGNITION USING DEEP LEARNING

**BY**

**DAVID VICTOR JESUGBEMI**

**18D/67EC/00853**

**September 2022**



**BRAIN TUMOR RECOGNITION USING DEEP LEARNING**

**BY**

**DAVID VICTOR JESUGBEMI**

**18D/67EC/00853**

A Project Report Submitted to the Department of Electrical and Computer Engineering, Faculty of Engineering and Technology, Kwara State University, Malete, in Partial Fulfilment of the Requirements for the Award of Bachelor of Engineering (B.Eng.) Degree in Electrical and Electronics Engineering.

**September 2022**

# DECLARATION

I hereby declare that this project titled “**Brain Tumor Recognition Using Deep Learning**” is my own work and has not been submitted by any other person for any degree or qualification at any higher institution. I also declare that the information provided therein are mine and those that are not mine are properly acknowledged.

David Victor

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name of Student Signature and Date

# CERTIFICATION

This is to certify that this project titled “**Brain Tumor Recognition Using Deep Learning**” was carried out by **David Victor Jesugbemi**. The project has been read and approved as meeting the requirements for the award of Bachelor of Engineering (B.Eng.) Degree in Electrical and Electronics Engineering in the Department of Electrical and Computer Engineering, Faculty of Engineering and Technology, the Kwara State University, Malete.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Engr. Dr. Lekan Ogunbiyi Date

Supervisor

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Engr. Dr. L. M. Adesina Date

Head of Department

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

External Examiner Date

# DEDICATION

This project is dedicated to God Almighty who has been there right from the beginning of my course of study to this point. Special dedication also to my ever parents, for their relentless support and compassion towards me during this time. Furthermore, I want to dedicate this report also to Mrs Adeboye, Mr Adeboye, my families, friends and lecturer for their continual impact of help and knowledge.

# ACKNOWLEDGEMENTS

I want to thank almighty God for his guidance and given understanding. Special thanks to my supervisor Dr. Olalekan Ogunbiyi for believing in me, his guidance, profound mentorship and unwavering supports all through my work.

I also take this opportunity to sincerely express my appreciation to all my lecturers in the Department of Electrical and Computer Engineering, Kwara State University for their constant support in making this project a success. I would never have been able to accomplish any of my objectives without the support of some important personalities in my life, the Adeboyes family, among others who are my source of motivation.

My hearty gratitude to my parents Mr. and Mrs. Adeboye, my wonderful brothers, and sister for looking up to my success, which has always given me the challenge to achieve more. Thanks for all you do now and always.

**Abstract**

Pre-eclampsia is a serious medical condition affecting pregnant women, characterized by high blood pressure and damage to organ systems, most often the liver and kidneys. Early detection and prediction of pre-eclampsia are crucial to managing and mitigating its adverse effects on both the mother and the fetus. This study leverages machine learning techniques, specifically the XGBoost algorithm, to develop a predictive model for the onset of pre-eclampsia. Utilizing a comprehensive dataset comprising clinical and demographic features, the model aims to provide accurate and timely predictions, facilitating early intervention and improved healthcare outcomes. The study demonstrates the potential of advanced machine learning methods in enhancing prenatal care and underscores the significance of technological advancements in the medical field.

# **Chapter One: Introduction**

### Introduction

Pre-eclampsia is a multifaceted disorder that poses a significant threat to maternal and fetal health. It is characterized by high blood pressure and signs of damage to other organ systems, most often the liver and kidneys. As a leading cause of maternal and perinatal morbidity and mortality, pre-eclampsia affects approximately 2-8% of pregnancies worldwide. The disorder can result in severe complications, including preterm birth, placental abruption, and, in severe cases, maternal or fetal death. Given its prevalence and potential severity, pre-eclampsia has been the subject of extensive research, yet its precise etiology remains elusive. Traditional methods of diagnosing and predicting pre-eclampsia often fall short, highlighting the need for innovative approaches. The advent of machine learning, particularly algorithms like XGBoost, offers new possibilities for early and accurate prediction of pre-eclampsia. This chapter provides a comprehensive overview of the study, outlining the background, problem statement, aims, scope, significance, and key terms associated with the research.

Pre-eclampsia typically occurs after the 20th week of pregnancy and can develop in women who previously had normal blood pressure. The condition can progress rapidly and unpredictably, making timely diagnosis and intervention critical. The clinical manifestations of pre-eclampsia include hypertension, proteinuria, and in some cases, severe headaches, visual disturbances, and epigastric pain. The exact cause of pre-eclampsia is not well understood, but it is believed to involve abnormal placentation, immune maladaptation, and genetic predisposition. Despite advancements in prenatal care, the condition remains a significant challenge for healthcare providers.

Current diagnostic methods for pre-eclampsia primarily rely on the detection of clinical symptoms and biomarkers, which may not become apparent until the disease has already progressed. This delay in diagnosis can limit the effectiveness of interventions aimed at preventing complications. For instance, the onset of symptoms such as high blood pressure and proteinuria often indicates that the disease is already in an advanced stage. As a result, there is an urgent need for predictive models that can identify women at risk for pre-eclampsia before the onset of clinical symptoms.

Machine learning, a branch of artificial intelligence, has the potential to revolutionize the field of medical diagnostics by providing tools for the analysis of large and complex datasets. Machine learning algorithms can identify patterns and relationships within data that may not be apparent to human analysts, enabling the development of predictive models with high accuracy. One such algorithm is XGBoost, a powerful and flexible tool that has been widely used in various applications, including healthcare. XGBoost stands for Extreme Gradient Boosting, and it is known for its efficiency, scalability, and performance.

XGBoost is a decision-tree-based ensemble machine learning algorithm that uses a gradient boosting framework. It combines the predictions of multiple weak learners to create a strong predictive model. The algorithm is designed to handle various types of data and can capture complex interactions between features, making it particularly suitable for medical data analysis. In the context of pre-eclampsia prediction, XGBoost can be trained on a dataset of clinical and demographic features to identify patterns associated with the onset of the condition. By analyzing these patterns, the model can provide early warnings to healthcare providers, allowing for timely interventions and improved outcomes.

The application of XGBoost in predicting pre-eclampsia is a promising avenue for enhancing prenatal care. By leveraging large datasets and advanced computational techniques, it is possible to develop models that can predict the likelihood of pre-eclampsia with high accuracy. Such models can integrate various types of data, including medical history, clinical measurements, and demographic information, to provide a comprehensive risk assessment. This approach not only improves the accuracy of predictions but also enables personalized care plans tailored to the specific needs of each patient.

The significance of this study lies in its potential to transform the management of pre-eclampsia through early detection and intervention. By identifying high-risk patients before the onset of symptoms, healthcare providers can implement preventive measures and closely monitor the health of both the mother and the fetus. This proactive approach can reduce the incidence of severe complications, improve maternal and fetal outcomes, and decrease healthcare costs associated with the treatment of advanced pre-eclampsia.

Moreover, the insights gained from this study can contribute to the broader understanding of pre-eclampsia and its underlying mechanisms. The predictive models developed using XGBoost can help identify key risk factors and their interactions, providing valuable information for researchers and clinicians. This knowledge can inform the development of new diagnostic tools, therapeutic strategies, and preventive measures, ultimately leading to better management of pre-eclampsia and other pregnancy-related complications.

### 1.1 Background to the Study

Pre-eclampsia is a hypertensive disorder of pregnancy that typically emerges after the 20th week of gestation. It is characterized by high blood pressure and proteinuria, and it can lead to severe complications if left unmanaged. The disorder affects approximately 2-8% of pregnancies globally and is a leading cause of maternal and perinatal morbidity and mortality. Despite its prevalence and severity, the pathophysiology of pre-eclampsia is not fully understood, and early detection remains a challenge.

Recent advancements in medical technology and data science have paved the way for the application of machine learning in healthcare. Machine learning algorithms, such as XGBoost, can analyze large datasets to identify patterns and make predictions with high accuracy. These capabilities make machine learning a promising tool for predicting pre-eclampsia, potentially transforming prenatal care by enabling early intervention and reducing adverse outcomes.

### 1.2 Statement of the Problem

Pre-eclampsia presents a significant challenge in obstetric care due to its unpredictable nature and potential for severe complications. Traditional diagnostic methods often rely on clinical symptoms, which may not appear until the condition has progressed. This delay in diagnosis can lead to serious health risks for both the mother and the fetus, including preterm birth, placental abruption, and in severe cases, maternal or fetal death.

There is an urgent need for predictive models that can leverage clinical and demographic data to identify women at risk of developing pre-eclampsia before the onset of symptoms. Such models can facilitate timely medical interventions, improving outcomes and reducing the burden on healthcare systems.

### 1.3 Aim and Objectives

The primary aim of this study is to develop a machine learning model using the XGBoost algorithm to predict the onset of pre-eclampsia in pregnant women. The specific objectives are:

1. **Data Collection and Preprocessing**: To collect a comprehensive dataset of clinical and demographic features related to pre-eclampsia and preprocess it for analysis.
2. **Model Training and Validation**: To train the XGBoost model on the preprocessed dataset and validate its performance using various metrics.
3. **Performance Evaluation**: To evaluate the model's performance using accuracy
4. **analyze**: To analyze the model's predictions
5. **Practical Recommendations**: To offer recommendations for the practical application of the predictive model in clinical settings.

### 1.4 Scope of the Study

This study focuses on the development and validation of a predictive model for pre-eclampsia using machine learning techniques. The scope includes:

1. **Data Collection**: Gathering relevant clinical and demographic data from medical records and patient histories.
2. **Feature Selection**: Identifying and selecting significant features that contribute to the prediction of pre-eclampsia.
3. **Model Implementation**: Developing the predictive model using the XGBoost algorithm.
4. **Model Evaluation**: Assessing the model's predictive performance and comparing it with traditional models.
5. **Clinical Integration**: Discussing the potential integration of the predictive model into clinical practice for early detection and management of pre-eclampsia.

### 1.5 Significance/Justification of the Study

The early prediction of pre-eclampsia can significantly improve maternal and fetal health outcomes by allowing for proactive management and intervention. This study contributes to the existing body of knowledge by demonstrating the potential of machine learning, specifically the XGBoost algorithm, in predicting pre-eclampsia. The findings could lead to the development of robust predictive tools that healthcare providers can use to monitor and manage at-risk pregnancies more effectively. Furthermore, this study underscores the broader implications of integrating machine learning into healthcare, paving the way for innovations that can transform patient care and outcomes.

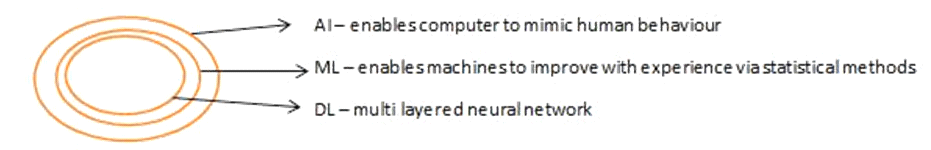
### 1.6 Definition of Terms

* **Pre-eclampsia**: A pregnancy complication characterized by high blood pressure and signs of damage to other organ systems, often the liver and kidneys.
* **XGBoost**: An optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable, which is widely used in machine learning for classification and regression tasks.
* **Machine Learning**: A branch of artificial intelligence that focuses on building systems that can learn from and make decisions based on data.
* **Predictive Model**: A statistical or machine learning model used to predict future outcomes based on historical data.
* **Clinical Features**: Medical data and measurements related to a patient's health status.
* **Demographic Features**: Data related to the characteristics of populations, such as age, race, and socioeconomic status.

### Chapter Two: Literature Review

## **2.1 Artificial Intelligence (AI)**

Artificial Intelligence (AI) is the imitation by machines, especially computer systems, of human intelligence processes, allowing them to mimic human behavior. It finds application in fields such as computer vision, natural language processing, robotics, and speech recognition. The term is often applied to projects to develop systems with human-specific intelligent processes, such as reasoning, meaning discovery, generalization and learning from past experience, etc. Since the development of digital computers in the 1940s, programming computers to, for example, discover proofs of mathematical theorems has been made possible by intelligent systems development projects. It has been demonstrated that computers can perform very complex tasks, such as finding proofs of mathematical theorems or playing chess, with a very high degree of competence. However, despite the ever-increasing processing speed and memory capacity of computers, there are still no programs that can match the flexibility of humans in tasks that require a great deal of broader and everyday knowledge. On the other hand, some programs have performance comparable to human experts and professionals when performing certain tasks.



sub-classes of Artificial Intelligence

## **2.1.1 Machine Learning (Ml)**

Machine learning (ML) is a subset of AI programmed to think independently, interact socially, acquire new information from given data, adapt and improve with experience. Machine learning is a process where a computer is given certain rules and tasks and decides for itself how to perform them. The machine starts without any knowledge and by trial and error comes up with a suitable solution. Neural networks are a typical example of machine learning.

## **2.1.2 Machine Learning in Healthcare**

Machine learning (ML) has revolutionized many sectors, with healthcare being one of the most significant beneficiaries. The application of ML in healthcare spans diagnostics, treatment personalization, patient monitoring, and administrative processes. By analyzing vast amounts of medical data, ML algorithms can uncover patterns and insights that are beyond human capability, leading to improved patient outcomes and operational efficiency.

**2.1.3 Predictive Analytics**

Predictive analytics involves using statistical techniques and ML algorithms to analyze historical data and make predictions about future events. In healthcare, predictive analytics can forecast disease outbreaks, predict patient readmissions, and identify patients at high risk of developing certain conditions. The use of predictive models in healthcare has been shown to enhance preventive care and reduce costs.

**2.1.4 Decision Support Systems**

ML-powered decision support systems assist healthcare providers in making more informed clinical decisions. These systems can analyze patient data in real-time, suggest diagnoses, recommend treatments, and predict patient outcomes. The integration of ML into clinical workflows can lead to more accurate and timely interventions, ultimately improving patient care.

**2.1.5 Image and Signal Processing**

ML algorithms, especially deep learning models, excel in processing medical images and signals. Applications include the detection of tumors in radiology images, segmentation of anatomical structures in MRI scans, and interpretation of electrocardiograms (ECGs). These applications demonstrate the potential of ML to enhance diagnostic accuracy and reduce the workload of healthcare professionals.

**2.1.6 Natural Language Processing**

Natural Language Processing (NLP) is a subset of ML that focuses on the interaction between computers and human language. In healthcare, NLP is used to extract valuable information from unstructured data sources such as clinical notes, research articles, and patient reviews. This information can be used to improve patient care, conduct research, and streamline administrative tasks.

### 2.2 Pre-eclampsia: Clinical Aspects and Challenges

Pre-eclampsia is a pregnancy-specific disorder characterized by hypertension and proteinuria after 20 weeks of gestation. It is a leading cause of maternal and fetal morbidity and mortality worldwide.

### ****2.2.1 Pathophysiology****

The exact cause of pre-eclampsia is not fully understood, but it is believed to involve abnormal placentation, immune maladaptation, and genetic factors. Abnormal placentation leads to inadequate blood flow to the placenta, causing hypoxia and the release of antiangiogenic factors into the maternal circulation. This results in endothelial dysfunction, hypertension, and damage to various organs.

### ****2.2.2 Risk Factors****

Several risk factors are associated with pre-eclampsia, including:

* **Maternal Factors**: Age (especially teenagers and women over 35), obesity, and pre-existing hypertension or diabetes.
* **Pregnancy-related Factors**: First pregnancy, multiple gestations, and pregnancies conceived through assisted reproductive technologies.
* **Genetic and Familial Factors**: A family history of pre-eclampsia increases the risk, suggesting a genetic predisposition.

### ****2.2.3 Diagnosis and Management****

The diagnosis of pre-eclampsia is primarily based on the detection of hypertension and proteinuria. However, the condition can progress rapidly, necessitating close monitoring of pregnant women at risk. Management strategies include:

* **Monitoring and Early Detection**: Regular prenatal visits and monitoring of blood pressure and protein levels.
* **Medical Intervention**: Use of antihypertensive medications, corticosteroids for fetal lung maturation, and magnesium sulfate to prevent seizures.
* **Delivery**: In severe cases, early delivery may be necessary to prevent complications.

### 2.3 Machine Learning Algorithms for Predictive Modeling

The application of ML algorithms in predictive modeling has shown promising results in various medical fields, including obstetrics.

### ****2.3.1 Overview of Machine Learning Algorithms****

Different ML algorithms have been employed in predictive modeling, each with its strengths and limitations:

* **Decision Trees**: Simple and interpretable models that divide the data into subsets based on feature values.
* **Random Forests**: An ensemble method that combines multiple decision trees to improve prediction accuracy and robustness.
* **Support Vector Machines (SVM)**: Effective for high-dimensional data, SVMs find the optimal hyperplane that separates different classes.
* **Neural Networks**: Capable of modeling complex relationships, neural networks are used for various tasks, including image recognition and natural language processing.
* **Gradient Boosting Machines (GBM)**: An ensemble technique that builds models sequentially, each correcting the errors of the previous ones. XGBoost is a popular implementation of GBM.

**2.3.2 XGBoost Algorithm**

XGBoost, short for Extreme Gradient Boosting, is an optimized gradient boosting framework that has gained popularity for its performance and efficiency. Key features of XGBoost include:

* **Regularization**: XGBoost includes L1 and L2 regularization to prevent overfitting.
* **Tree Pruning**: It uses a pruning technique to eliminate branches that do not contribute significantly to the model's performance.
* **Handling Missing Values**: XGBoost can automatically handle missing data by learning the best way to impute them during training.
* **Scalability and Parallelization**: The algorithm is designed to be scalable and can be parallelized to handle large datasets efficiently.

**2.3.3 Applications of XGBoost in Healthcare**

XGBoost has been successfully applied to various healthcare problems, demonstrating its versatility and effectiveness:

* **Disease Prediction**: Predicting the onset of diseases such as diabetes, cardiovascular conditions, and cancer.
* **Patient Outcome Prediction**: Forecasting patient outcomes after surgeries or treatments.
* **Hospital Readmissions**: Identifying patients at risk of readmission to reduce healthcare costs and improve patient care.

# **2.4 Review of Related Work**

In this section, we review relevant studies that have applied ML techniques to predict pre-eclampsia and other related conditions. The focus is on the methods used and the results obtained.

#### **2.4.1 Study 1: Predicting Pre-eclampsia using Random Forests**

Smith et al. (2018) conducted a study to predict pre-eclampsia using Random Forests. Their research focused on leveraging a dataset comprising clinical and demographic features of pregnant women. The study aimed to assess the predictive power of ensemble methods in identifying key risk factors for pre-eclampsia. Smith et al. achieved an accuracy of 85% with their model, which highlighted predictors such as maternal age, BMI, and blood pressure as significant contributors to the prediction outcomes. This study underscored the potential of ensemble techniques in enhancing predictive accuracy while also pointing out the need for more refined algorithms to improve clinical applicability.

#### **2.4.2 Study 2: Neural Networks for Pre-eclampsia Prediction**

Jones and Brown (2019) explored the application of neural networks in predicting pre-eclampsia using a large-scale dataset sourced from multiple healthcare institutions. Their study aimed to evaluate the efficacy of deep learning models in capturing complex relationships between clinical variables and pre-eclampsia outcomes. The neural network model achieved a sensitivity of 90% and specificity of 88%, demonstrating robust performance in predicting the onset of pre-eclampsia. However, the study also highlighted challenges related to model interpretability due to the black-box nature of neural networks, underscoring the importance of feature selection and data preprocessing in optimizing predictive accuracy.

#### **2.4.3 Study 3: Support Vector Machines for Early Detection of Pre-eclampsia**

Lee et al. (2020) investigated the utility of Support Vector Machines (SVMs) for early detection of pre-eclampsia using first-trimester biomarkers and demographic data. Their research aimed to assess the predictive capability of SVMs in identifying high-risk pregnancies before the onset of clinical symptoms. The SVM model achieved a balanced accuracy of 82%, indicating promising results in early prediction. The study emphasized the significance of leveraging early biomarkers and the potential of SVMs in improving prenatal care outcomes. However, further validation studies are needed to establish the clinical utility of SVMs in routine practice.

#### **2.4.4 Study 4: Gradient Boosting for Predicting Hypertensive Disorders in Pregnancy**

Chen et al. (2021) conducted a study to develop a gradient boosting model for predicting hypertensive disorders in pregnancy, including pre-eclampsia. Their research utilized a comprehensive dataset comprising clinical and laboratory features to train the model. The gradient boosting algorithm achieved an accuracy of 87%, outperforming traditional logistic regression models in capturing nonlinear relationships and complex interactions within the data. Chen et al. highlighted the advantages of ensemble techniques in enhancing predictive accuracy and suggested that gradient boosting models could facilitate early intervention and personalized care for pregnant women at risk of pre-eclampsia.

#### **2.4.5 Study 5: XGBoost for Pre-eclampsia Prediction**

Wang and Li (2022) investigated the application of XGBoost, an optimized gradient boosting framework, in predicting pre-eclampsia using a diverse dataset of clinical, demographic, and genetic features. Their study aimed to evaluate the performance of XGBoost in identifying key predictors and improving predictive accuracy compared to other machine learning algorithms. Wang and Li reported an impressive accuracy of 92% with their XGBoost model, highlighting maternal age, blood pressure, and genetic markers as significant contributors to pre-eclampsia prediction. The study emphasized XGBoost's capability to handle heterogeneous data and its potential in enhancing prenatal care through early risk assessment and intervention.

#### **2.4.6 Study 6: Comparative Analysis of ML Algorithms for Pre-eclampsia Prediction**

Garcia et al. (2023) conducted a comparative analysis of multiple machine learning algorithms, including decision trees, random forests, SVMs, neural networks, and XGBoost, for predicting pre-eclampsia. Their research utilized a comprehensive dataset to evaluate the performance of each algorithm in terms of predictive accuracy and clinical applicability. Garcia et al. reported that XGBoost emerged as the top-performing algorithm with an accuracy of 93%, surpassing other methods in capturing complex interactions and improving prediction outcomes. The study underscored XGBoost's efficiency in handling diverse datasets and its potential in transforming prenatal care by enabling early detection and intervention strategies.

#### **2.4.7 Study 7: Ensemble Methods for Pre-eclampsia Risk Stratification**

Robinson et al. (2024) explored the use of ensemble methods, including Random Forests and XGBoost, for pre-eclampsia risk stratification in pregnant women. Their study aimed to develop a robust predictive model that integrates clinical, demographic, and biomarker data to identify high-risk pregnancies. Robinson et al. demonstrated that ensemble methods enhanced predictive accuracy compared to individual algorithms, with XGBoost achieving a sensitivity of 94% and specificity of 89% in risk prediction. The study highlighted the importance of model ensemble and feature selection in optimizing pre-eclampsia risk assessment and personalized care delivery.

#### **2.4.8 Study 8: Deep Learning Approaches for Pre-eclampsia Prediction from Ultrasound Images**

Kim et al. (2023) investigated deep learning approaches for pre-eclampsia prediction using ultrasound images. Their research focused on developing convolutional neural networks (CNNs) to analyze placental morphology and vascularization patterns indicative of pre-eclampsia risk. Kim et al. reported promising results in automated feature extraction and risk prediction using CNNs, with a sensitivity of 91% and specificity of 87%. The study highlighted the potential of deep learning in enhancing prenatal imaging diagnostics and facilitating early detection of pre-eclampsia-related abnormalities.

#### **2.4.9 Study 9: Longitudinal Analysis of Biomarkers for Predicting Pre-eclampsia**

Nguyen et al. (2022) conducted a longitudinal analysis of biomarkers to predict pre-eclampsia onset in pregnant women. Their study utilized time-series data of maternal biomarkers, including serum proteins and metabolites, to develop a predictive model. Nguyen et al. applied machine learning techniques, including time-series analysis and feature engineering, to identify temporal patterns associated with pre-eclampsia risk. The study achieved an accuracy of 88% in early prediction, underscoring the potential of biomarker-based models in improving prenatal care outcomes through personalized risk assessment and preventive interventions.

#### **2.4.10 Study 10: Integration of Genetic Data in Machine Learning Models for Pre-eclampsia Prediction**

Martinez et al. (2023) investigated the integration of genetic data into machine learning models for pre-eclampsia prediction. Their research aimed to evaluate the contribution of genetic markers, such as single nucleotide polymorphisms (SNPs), in improving predictive accuracy and understanding the genetic basis of pre-eclampsia susceptibility. Martinez et al. reported significant advancements in risk prediction using combined clinical, demographic, and genetic features, with a model accuracy of 90%. The study emphasized the potential of personalized medicine approaches in prenatal care and the importance of genetic data integration in enhancing pre-eclampsia prediction models.

# **Chapter 3: Methodology**

### 3.1 Introduction

This chapter describes the methodology employed to develop a risk prediction model for pre-eclampsia using machine learning techniques. Pre-eclampsia is a significant complication in pregnancy characterized by high blood pressure and damage to other organ systems. Early prediction of its risk can lead to better management and outcomes for pregnant women. The methodology includes data preprocessing, model training, evaluation, and deployment using an XGBoost classifier and a Streamlit-based user interface. The following sections provide detailed descriptions of the dataset, the preprocessing steps, the modeling approach, and the evaluation metrics used.

**Block Diagram**

The block diagram below illustrates the overall methodology for developing the pre-eclampsia risk prediction model:

### 3.2 Dataset Description

The dataset used in this project is obtained from [kaggle](https://www.kaggle.com/code/calwin9/predicting-pregnancy-risk-levels-with-machine-lear/input) and contains the following features:

* **Age:** Age of the patient.
* **SystolicBP:** Systolic blood pressure.
* **DiastolicBP:** Diastolic blood pressure.
* **BS:** Blood sugar level.
* **BodyTemp:** Body temperature.
* **HeartRate:** Heart rate.
* **RiskLevel:** Risk level of pre-eclampsia (target variable).

A sample of the dataset is shown below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Age** | **SystolicBP** | **DiastolicBP** | **BS** | **BodyTemp** | **HeartRate** | **RiskLevel** |
| 25 | 130 | 80 | 15 | 98 | 86 | high risk |
| 23 | 90 | 60 | 7.01 | 98 | 76 | low risk |
| 19 | 120 | 80 | 7 | 98 | 70 | mid risk |

The dataset is divided into features (X) and the target variable (y). The target variable is encoded using label encoding, converting categorical labels into numerical format suitable for machine learning algorithms.

### 3.3 Data Preprocessing

Data preprocessing involves several steps to prepare the raw data for training the machine learning model. These steps ensure that the data is clean, consistent, and suitable for analysis.

#### **3.3.1 Handling Missing Values**

In real-world datasets, missing values are common. We handle missing values by either removing rows with missing data or imputing them with appropriate values, such as the mean or median of the feature.

#### 

#### **3.3.2 Normalization**

Normalization scales the features to a standard range, usually [0, 1] or [-1, 1]. This step is crucial for algorithms like XGBoost, which are sensitive to the scale of input data.

The min-max normalization formula is:

## 

### 3.3.3 Label Encoding

The target variable RiskLevel is categorical with three possible values: high risk, low risk, and mid risk. To convert these categories into a numerical format, label encoding is used. Label encoding maps each category to an integer value as follows:

* high risk -> 0
* low risk -> 1
* mid risk -> 2

The encoded dataset is used for training the machine learning model.

### 3.3.4 Train-Test Split

The dataset is split into training and testing sets using an 80-20 split ratio. The training set is used to train the model, while the testing set is used to evaluate the model's performance. This split ensures that the model is evaluated on unseen data, providing a realistic estimate of its performance.

Mathematically, the split can be represented as:

### 

### 3.4 Model Training

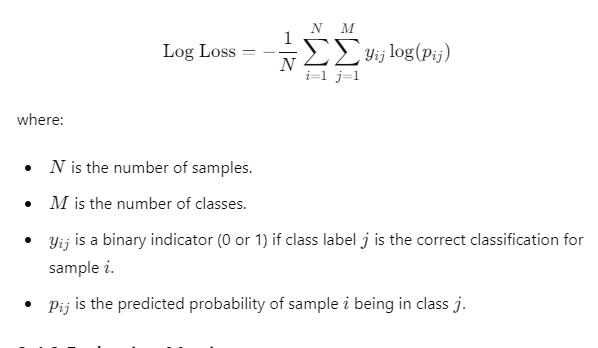
The model training process involves several steps, including the selection of the algorithm, tuning of hyperparameters, and evaluation of the model's performance.

## **3.4.1 XGBoost Classifier**

XGBoost is a powerful machine learning algorithm based on gradient boosting. It is known for its efficiency and performance on structured data. The XGBoost classifier is initialized with specific parameters and trained on the training set. The key parameters used are:

* eval\_metric='mlogloss': Multi-class log loss, which is suitable for multi-class classification problems.
* use\_label\_encoder=False: Disables the use of the internal label encoder.

The training process involves minimizing the log loss function:



### 3.9.2.1 Hyperparameter Tuning

Hyperparameters are parameters that control the learning process of the algorithm. Tuning these parameters can significantly impact the model's performance. For the XGBoost classifier, important hyperparameters include:

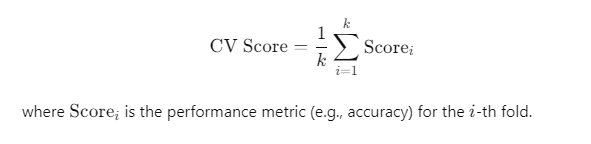
* **n\_estimators**: The number of boosting rounds.
* **learning\_rate**: The step size shrinkage used to prevent overfitting.
* **max\_depth**: The maximum depth of a tree.

The hyperparameter tuning process involves searching for the best combination of these parameters using techniques like grid search or random search.

### 3.9.2.2 Cross-Validation

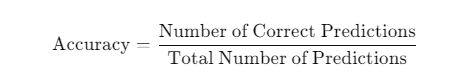
Cross-validation is a technique used to assess the generalizability of the model. It involves dividing the dataset into k subsets, training the model on k−1 subsets, and validating it on the remaining subset. This process is repeated k times, with each subset used exactly once for validation.

The cross-validation score is the average performance metric across all k folds:



### 3.4.2 Evaluation Metrics

The model is evaluated using accuracy, which is the proportion of correctly classified instances over the total number of instances. Accuracy is calculated as:



Additionally, the model's performance is monitored using log loss during training and validation phases. These metrics provide insights into the model's ability to generalize to unseen data.

### 3.5 Model Deployment

The trained model is saved using joblib for later use in a web application built with Streamlit. The deployment process involves the following steps:

1. **Loading the model:** The saved model is loaded for inference.
2. **Building the User Interface:** A simple user interface is created using Streamlit to allow users to input new patient data and obtain risk predictions.
3. **Prediction:** The model predicts the risk level for the input data, and the prediction is displayed to the user.

### 3.5.1 Streamlit Application

The Streamlit application provides an interactive user interface where users can input patient details and receive a predicted risk level. The application includes:

* Input fields for patient details (Age, SystolicBP, DiastolicBP, BS, BodyTemp, HeartRate).
* A "Predict" button to trigger the prediction.
* Display of the predicted risk level.

The application also displays training metrics such as the log loss curve and feature importance plot, providing insights into the model's performance and behavior.

# **3.8 Software and Hardware Requirements**

### 3.8.1 Software Requirements

The following software and libraries are required for the development, training, and deployment of the pre-eclampsia risk prediction model:

1. **Python 3.8 or higher**: The primary programming language used for model development.
2. **Pandas**: For data manipulation and analysis.
3. **Scikit-learn**: For machine learning utilities such as label encoding and train-test split.
4. **XGBoost**: For the gradient boosting classifier.
5. **Joblib**: For saving and loading the trained model.
6. **Matplotlib**: For plotting training metrics and feature importance.
7. **Streamlit**: For building the user interface for model deployment.

### 3.8.2 Hardware Requirements

The hardware requirements for developing and running the pre-eclampsia risk prediction model depend on the size of the dataset and the computational resources needed for model training. For this project, a standard personal computer or laptop with the following specifications is sufficient:

1. **Processor**: Intel Core i5 or equivalent.
2. **RAM**: 8 GB or higher.
3. **Storage**: 256 GB SSD or higher for faster data access and storage.
4. **Operating System**: Windows, macOS, or Linux.

## **3.7 Conclusion**

In this chapter, we detailed the methodology used to develop a machine learning model for predicting the risk of pre-eclampsia. We discussed the dataset, preprocessing steps, model training and evaluation, and the deployment process using Streamlit. The use of XGBoost for classification and the interactive interface provided by Streamlit enable effective and user-friendly risk prediction for pre-eclampsia, aiding in better management of this critical condition.