

Software Engineering Department

Braude College

Capstone Project Phase A

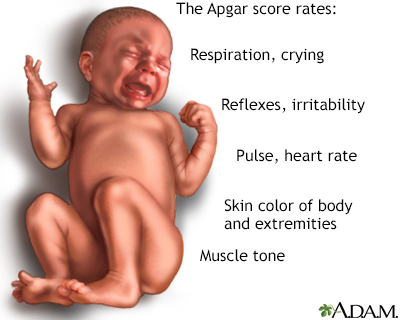
**Enhancing Perinatal Care with AI-Driven Apgar Score Prediction**

Project code: 25-2-R-4

Supervisors: Gur Arye Yehuda

Yael Kanter 315823245

Ofir Ofek 312168347

Git: <https://github.com/yaelkanter/Apgar-Score-Prediction>

**Table of Contents**

### 

1. Abstract
2. Introduction
3. Literature Review  
    3.1 Predicting Apgar Scores Using Machine Learning  
    3.2 Challenges and Limitations  
    3.3 Future Opportunities
4. Background  
    4.1 SMOT
5. Related Studies and Contributions
6. Research Process  
    6.1 Data Collection and Preprocessing  
    6.2 Data Cleaning Strategy  
    6.3 Key Data Challenges and Solutions  
    6.4 Model Selection and Training  
    6.5 Evaluation and Testing

6.6 Key Predictive Insights

1. Future Work

7.1 Using More Data and Richer Information

7.2 Focusing on Specific Risks

7.3 Trying Smarter Algorithms

7.4 Making It Work in Real Time

7.5 Testing It in Real Hospitals

7.6 Thinking About Ethics and Privacy

8. References

**1.Abstract**

As the health of mothers and babies gets more attention around the world, there’s a growing need for smart tools to help doctors make better choices during childbirth. One of the main ways to check a baby’s condition after birth is the Apgar score, which looks at things like heart rate and breathing. While this score is helpful, doctors don’t yet have a good way to predict it before birth, which can make it harder to prepare for problems.

In this project, we use Artificial Intelligence (AI) to build a model that can predict a baby’s Apgar score before birth. We train machine learning algorithms using real medical data, like the mother’s health, details about the pregnancy, and labor information. The model learns to find patterns and make accurate predictions.

Our goal is to develop a machine learning model that can provide quick insights about the baby’s expected condition based on patient data.  
 This can help spot risky births earlier, plan better care, and use medical resources more effectively. We believe AI can make childbirth safer and more personalized for both mothers and babies

**Keywords:** Perinatal care · Apgar score · Artificial Intelligence · Machine Learning · Predictive modeling · Childbirth · Maternal health · Neonatal outcomes

**2.** **Introduction**

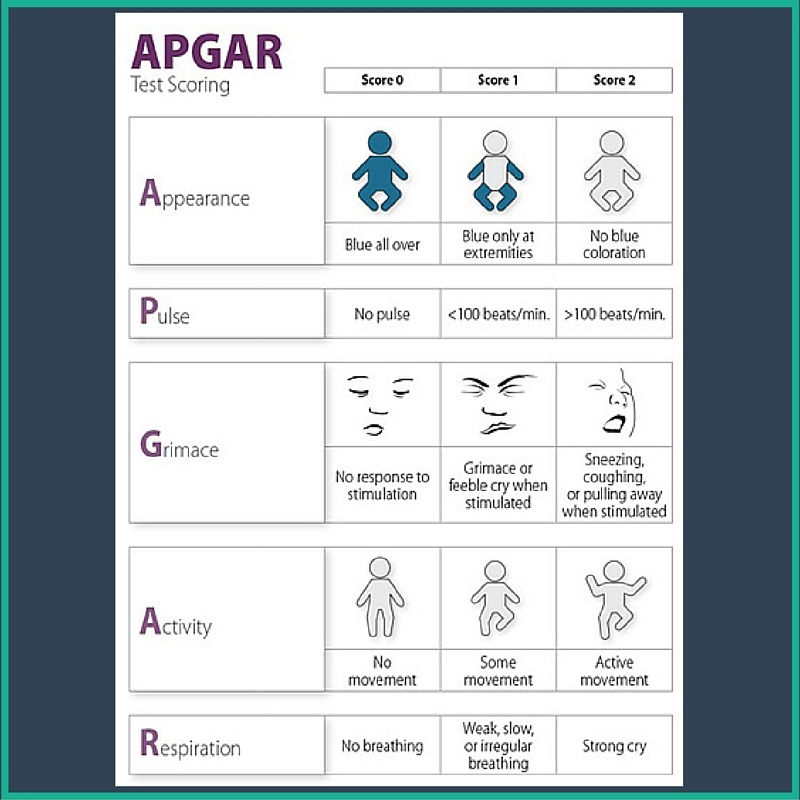
Perinatal care is very important for making childbirth safer and improving the health of both mothers and babies. One of the main tools used to check a newborn’s condition right after birth is the Apgar score, created by Dr. Virginia Apgar in 1952. This score looks at five things: heart rate, breathing, muscle tone, reflexes, and skin color, checked at one and five minutes after birth [1]. A low Apgar score is often linked to a higher risk of health problems or death in newborns [2].

Even though the Apgar score is very important, it can only be used after the baby is born. It doesn’t help doctors know in advance if there might be problems during birth. Right now, doctors mostly use their experience and basic monitoring tools, which are not always enough to predict how the baby will do. This can cause delays in giving the right care and may affect the baby’s health both right after birth and later in life.[7]

New progress in Artificial Intelligence (AI) and Machine Learning (ML) gives us new ways to improve care during and before birth. Studies have shown that ML models can look at complex data-like the mother’s health, pregnancy history, baby’s condition, and details about labor-to predict if a baby might get a low Apgar score. In one study, ML was used to analyze birth data and was able to accurately predict which babies would get a score below 7 at five minutes after birth, which is a sign of high risk.[4][5]

These new technologies make it possible to create helpful tools for doctors and nurses. By giving early warnings, these tools can help medical teams get ready before problems happen. They can also help hospitals use NICU beds more wisely and reduce stress and costs for both families and healthcare systems.

This project aims to develop a simple AI-powered tool-like an app or service-that doctors can use to predict a baby’s Apgar score before birth. By entering information about the mother, pregnancy, and labor, the system will quickly give a prediction to help doctors prepare in advance. The goal is to support medical teams with early insights, improve care, and shift from reacting to preventing complications.



**Figure 1:** APGAR test scoring chart, indicating points 0–2 for Appearance, Pulse, Grimace, Activity, and Respiration. *Source: Medicine Hack. (2010, May 4).* APGAR scoring – table, mnemonic [Infographic]. Retrieved from<https://www.medicinehack.com/2010/05/apgar-scoring.html>

**3. Literature Review**

Perinatal care has been a subject of increasing attention due to its critical role in reducing maternal and neonatal mortality. The Apgar score, introduced in 1952, remains one of the most widely used tools for assessing a newborn’s condition at birth. However, it serves as a retrospective evaluation tool and does not allow clinicians to anticipate complications prior to delivery [1].

**3.1 Predicting Apgar Scores Using Machine Learning**

Recent advances in Machine Learning (ML) and Artificial Intelligence (AI) have enabled researchers to develop models that predict neonatal outcomes using perinatal data. One such study by Mdoe et al. examined 7,716 induced vaginal deliveries in Tanzania. The researchers applied several ML techniques to predict low Apgar scores (<7) at five minutes and identified birth weight, maternal age, and gestational age as significant predictors. To address class imbalance in the dataset, they used SMOTE (Synthetic Minority Over-sampling Technique) to enhance model performance [7].

Similarly, Nourani et al. proposed a distributed big data analytics approach using a national birth dataset in the United States to predict five-minute Apgar scores. Their ensemble model demonstrated high prediction accuracy and identified fetal head circumference, fetal weight, and gestational week as key features [4]. The authors highlighted the value of using large, real-world datasets for developing robust predictive models in perinatal care.

Another important contribution was made by Rocha et al., who conducted a systematic review of AI-based models for predicting maternal and neonatal mortality. The review found that random forests, logistic regression, and support vector machines were commonly used, and key predictive features included gestational age, maternal age, and Apgar score itself [6].

**3.2 Challenges and Limitations**

Even though AI models show a lot of promise, there are still some big challenges. One of the main problems is class imbalance-low Apgar scores don’t happen very often, so the model doesn’t get enough examples to learn from. This can lead to biased results. Methods like SMOTE are often used to fix this, but they need to be tested carefully so the model doesn’t just memorize the data (called overfitting) [7].

Another important part of training is choosing the right number of epochs (training cycles). If we use too few, the model doesn’t learn enough-this is called underfitting. But if we use too many, the model may learn the training data too well and won’t work well on new data-this is overfitting. In our project, we used a method called early stopping and watched the validation loss to stop training at the right time.

Early stopping is a method that stops training when the model stops improving on validation data, to prevent overfitting. We usually found that training for 30 to 50 epochs gave good results, depending on how complex the model was. We also used dropout layers and regularization to help prevent overfitting.

Another challenge is that health data can be very different from one hospital to another. Some hospitals collect different information or might be missing important features. This makes it hard for one model to work well everywhere. To fix this, it’s important to make data more standardized and to use explainable AI (XAI) methods so doctors can understand and trust how the model makes its decisions [10].

**3.3 Future Opportunities**

Looking ahead, future research should focus on building AI models that are easy to understand and can be used in real hospital systems. Models that can show which features had the biggest impact on each prediction are more likely to be trusted and used by doctors. Also, using larger and more diverse datasets-from different hospitals and populations-will help the models work better in different places and reduce bias in the predictions [6].

**4.** **Background**

Childbirth is a life-changing event-both miraculous and risky. For healthcare providers, one of the biggest challenges is making sure both the mother and baby come through the experience safely. In these critical moments, doctors rely on tools like the Apgar score, which offers a quick assessment of a newborn’s condition in the first minutes of life. Developed by Dr. Virginia Apgar in 1952, this scoring system evaluates five key indicators-heart rate, respiratory effort, muscle tone, reflex response, and skin color-to determine whether a newborn requires immediate medical intervention [1].

While the Apgar score is essential in the minutes after birth, it doesn’t offer much help before delivery. Clinicians must still depend largely on subjective judgment, basic fetal monitoring, and experience to make decisions during labor. Unfortunately, these methods can fall short, especially in cases where complications arise quickly and unexpectedly [2]. This gap between assessment and anticipation creates a serious challenge: how can we better prepare for emergencies before they happen?

In recent years, the medical world has started turning to Artificial Intelligence (AI) and Machine Learning (ML) to help close this gap. These technologies excel at finding patterns in complex, multi-dimensional data-something that is nearly impossible to do manually. In perinatal care, several studies have already shown promising results. For example, Mdoe et al. trained machine learning models on birth data from Tanzania and found that key features like maternal age, gestational age, and birth weight could help predict low Apgar scores at five minutes with a high degree of accuracy [7]. Similarly, Nourani et al. used a national U.S. birth dataset and built an ensemble AI model that successfully predicted Apgar outcomes using features such as fetal head circumference and gestational week [4].

These results suggest that AI-based tools have the potential to transform how we approach childbirth, shifting from reactive care to proactive planning. If healthcare teams could identify at-risk births before labor even begins, they would be better equipped to act quickly, allocate resources like neonatal intensive care units, and tailor interventions to each case. This would not only improve clinical outcomes but also reduce the emotional and financial stress placed on families.

Still, challenges remain. One major challenge is data imbalance-low Apgar scores are relatively rare, which makes it difficult to train models without introducing bias. Researchers like Rocha et al. have emphasized the need for techniques like SMOTE (Synthetic Minority Over-sampling Technique) to address this issue, while also calling for more interpretable models that clinicians can trust and understand [5][6].

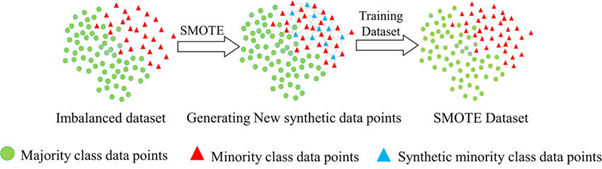
In this project, we propose developing an AI-driven predictive tool that can be used during labor or prenatal checkups. Clinicians will be able to enter available patient data-such as maternal history, fetal measurements, and labor parameters-and receive an immediate prediction of the baby’s expected Apgar score. This tool could become an invaluable part of the clinical workflow, supporting real-time decisions and ultimately making childbirth safer for everyone involved.

By combining cutting-edge technology with human-centered care, we aim to contribute a meaningful step forward in the future of personalized perinatal medicine.

**4.1** **SMOT**

To deal with the problem of class imbalance-where most babies have normal Apgar scores and only a few have low scores-we used a method called SMOTE (Synthetic Minority Over-sampling Technique). SMOTE helps balance the data by creating new, fake examples of the rare cases. It does this by looking at patterns in the real data and generating new samples that are similar, instead of just copying the rare cases (which could cause overfitting).

These new examples give the model more variety and help it understand what low Apgar scores cases can look like. By adding more of these rare cases to the training data, SMOTE helps the machine learning model learn in a more balanced and fair way. This is especially important in healthcare, where missing a baby at risk (a false negative) can be very dangerous. SMOTE improves the model’s ability to catch those rare but critical cases.



**Figure 2**: SMOTE algorithm generating synthetic minority samples by interpolating between existing Source:<https://pub.aimind.so/synthetic-minority-over-sampling-technique-smote-empowering-ai-through-imbalanced-data-handling-d86f4de32ea3>

In our project, SMOTE was an important part of preparing the data before training the prediction model. Without using this balancing method, the model would mostly learn to predict normal outcomes, because those are the most common in the data. SMOTE gave the model a better chance to recognize early signs of problems that could lead to low Apgar scores.

Past studies have shown how helpful this method can be. For example, Mdoe et al. used SMOTE in a large study of vaginal births in Tanzania, where low Apgar scores were rare but very serious [7]. Rocha et al. also pointed out that using techniques like SMOTE-together with models that doctors can understand-is very important for building clinical tools that can be trusted [6].

However, it's also important to remember that SMOTE creates new, synthetic data based on existing cases. In a medical setting, this can be risky because the new samples are not real and may not fully capture the complexity of real-life health conditions. That’s why we used SMOTE carefully and always checked that the results still made sense medically.

In the end, SMOTE is not just a technical step-it plays a big role in making sure that AI models in healthcare are fair, accurate, and able to catch rare but life-threatening cases, even when the real-world data is unbalanced.

**Algorithm Steps:**

1. **Identify minority class samples** The algorithm begins by locating all data points that belong to the minority class.
2. **Find k-nearest neighbors** For each minority class point, SMOTE identifies *k* of its nearest neighbors from the same class (commonly k = 5).
3. **Generate new samples** For each selected point, SMOTE randomly selects one of its neighbors and generates a synthetic sample along the line between the two.

This is done using the formula:

where is a random number between 0 and 1.

Also, if the original data has errors or strange cases (called outliers), SMOTE might accidentally create more of those, which could confuse the model instead of helping it.

So while SMOTE helps the model pay more attention to rare and important cases, we have to use it carefully. It’s important to check the results and make sure the model still makes sense in the real world.

### **5. Related Studies and Contributions**

Many recent studies have explored how machine learning (ML) and artificial intelligence (AI) can help predict newborn outcomes and improve care during pregnancy and birth. These studies show that computer models can find patterns in mother and baby data that doctors might miss using traditional methods.

**Apgar (1953)** [1] The Apgar score was first introduced by Dr. Virginia Apgar in 1953 as a quick way to check a newborn’s health. Her work became the base for modern newborn evaluations and is still widely used today. Our project builds on this idea by using AI to predict Apgar scores in advance.

**Casey et al. (2001)** [2] showed that low Apgar scores are strongly linked to higher chances of health problems or death in newborns. Their findings confirmed how important it is to predict Apgar scores early, which is why we chose it as our main target..

**Nourani et al. (2022)** [4] used a large birth dataset from the U.S. to build a strong prediction model using ensemble learning. They included data like the baby’s head size, weight, and pregnancy weeks, and reached high accuracy. Their work showed us the value of using large, varied datasets and helped us decide which features to include.

**Lee et al. (2023)** [5] used deep learning to predict early death and serious brain issues in premature babies. Their study showed how important it is to adjust model settings carefully (like learning rate and batch size) and to use metrics like recall when working with high-risk cases. This directly influenced how we trained and tested our models.

**Rocha et al. (2022)** [6] reviewed many AI models used to predict health outcomes for mothers and babies. They emphasized that models should be easy to understand and explained which algorithms are commonly used, such as logistic regression, support vector machines, and random forests. This helped us choose models that are both accurate and clear for doctors to interpret

**Tarimo et al. (2022)** [7] focused on predicting low Apgar scores in vaginal births at a hospital in northern Tanzania. They used machine learning to deal with real challenges in hospitals with limited resources. Like other studies, they used SMOTE to handle imbalanced data, which helped their model better identify babies at risk. Their study confirmed that machine learning can be useful in hospitals, even with limited data. They used similar features to ours-like maternal age, birth weight, and gestational age-and their success supported our approach, especially our decision to use SMOTE during preprocessing.

These studies showed us that our project is both important for healthcare and possible to build with the right tools. They guided us in choosing the right features, balancing the data, and testing our model in a way that makes sense for real hospitals.

**6. Research Process**

Our research process involves several key steps: data preprocessing, handling class imbalance using SMOTE, feature selection, and the application of machine learning (ML) algorithms. These steps are designed to extract meaningful patterns from perinatal health data to support accurate Apgar score prediction. However, the process is inherently dynamic and may evolve as the research progresses. This flexibility is essential to adapt to emerging challenges, such as data quality issues, model performance tuning, or discovering unexpected clinical correlations within the dataset:

**6.1 Data Collection and Preprocessing**

We used an anonymized dataset that included information about the mother's and baby’s health. Some of the features were the mother’s age, number of weeks of pregnancy, baby’s weight, length of labor, and type of delivery. Before training the machine learning models, we cleaned and normalized the data to make sure it was ready to use. Because most babies in the dataset had normal Apgar scores, the classes were unbalanced. To fix this, we used a method called SMOTE to create more examples of the rare cases and help the model learn from them better

### 

### **6.2 Data Cleaning Strategy**

To ensure the quality and reliability of our dataset before applying machine learning models, we followed a structured data cleaning pipeline. First, we addressed missing values by selecting only rows with complete numerical data using the **dropna()** method. This step helped reduce noise and inconsistencies that might interfere with model training.

Next, we applied the **Isolation Forest** algorithm to detect and eliminate statistical outliers. This technique analyzes the distribution of numerical features and identifies unusual records that may result from data entry errors or rare clinical anomalies. We kept only the most consistent 90% of records, improving data integrity without significantly reducing the dataset's diversity.

After filtering, we reloaded the cleaned dataset and replaced any remaining missing or undefined values with placeholder values (such as -1) to ensure compatibility with our models. This process allowed us to retain more rows while still flagging missing information for potential model learning.

This cleaning process not only improved data quality but also helped reduce overfitting risks by ensuring the model would not learn from erroneous or non-representative data.

**6.3** **Key Data Challenges and Solutions**

We used a large clinical dataset consisting of **1,149,433 birth records** and **24 features**. Each row represents a single birth event and includes detailed information about the mother, pregnancy, delivery process, and newborn outcomes.

The features in the dataset are grouped into three main categories:

#### **Maternal Information**

* **maternal\_age** – Mother’s age (numerical)
* **tp\_maternal\_education, tp\_maternal\_skin\_color** – Categorical data reflecting demographic and educational background

#### **Pregnancy and Delivery Details**

* **tp\_pregnancy** – Type of pregnancy ( single, multiple)
* **tp\_labor** – Type of delivery (vaginal, cesarean)
* **tp\_prenatal\_appointments –** Number of prenatal visits (numerical)
* **num\_gestations, num\_live\_births, num\_normal\_labors** – Pregnancy and delivery history

#### **Newborn and Clinical Outcomes**

* **newborn\_weight** – Baby’s birth weight in grams (numerical)
* **num\_gestational\_weeks** – Gestational age at birth (numerical)
* **cd\_apgar1, cd\_apgar5** – Apgar scores at 1 and 5 minutes after birth (ordinal)
* **neonatal\_death** – Binary target variable (0 = alive, 1 = neonatal death)

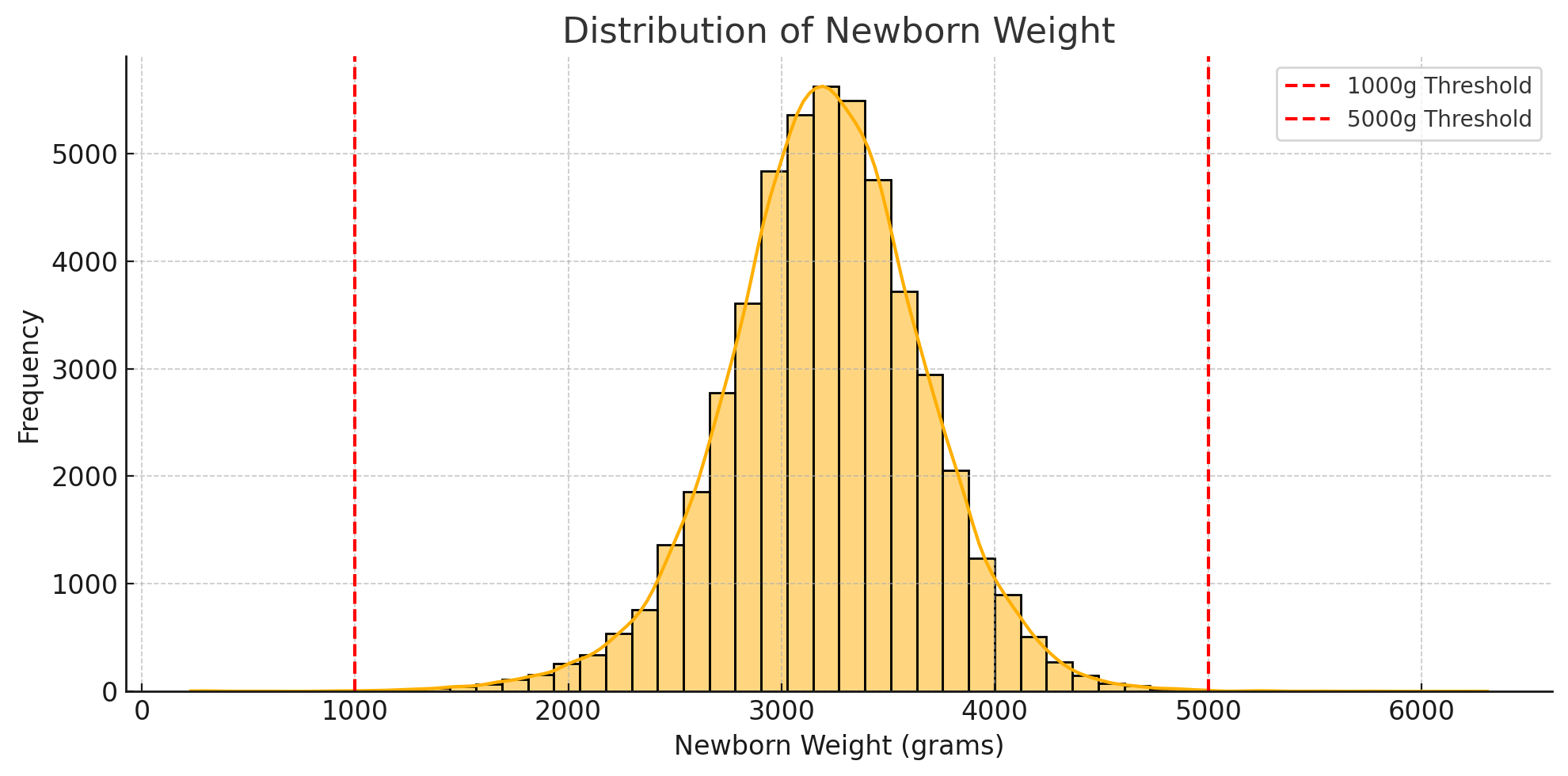
1. **Incomplete Data:**Clinical datasets often have missing or incomplete information. This can happen because of human mistakes, missing records, or equipment problems. These gaps can make it harder for the model to learn and can lead to unfair or incorrect predictions. To fix this, we used methods like filling in missing numbers with the average (mean) or using k-nearest neighbors (k-NN) to guess missing values based on similar data. For records that were missing a lot of information, we looked at them carefully to decide if we should keep them or remove them. This helped make sure our training data was still accurate and similar to real clinical situations.
2. **Class Imbalance:** A significant challenge in our dataset was the imbalance between the number of newborns with normal versus low Apgar scores. Since most infants are born healthy, the dataset was dominated by normal cases, making it difficult for the model to learn to recognize the rare, yet critical, low-score cases. To overcome this, we employed SMOTE (Synthetic Minority Over-sampling Technique), a method that generates synthetic examples of the minority class. This improved the model’s ability to detect high-risk births and reduced its bias toward predicting only common outcomes.

1. **Outliers and Anomalies:** Some data points showed extreme or inconsistent values, such as unusually high birth weights or implausible gestational ages. These outliers could be the result of data entry errors or rare medical conditions and had the potential to distort the training process. To ensure the model focused on realistic clinical patterns, we used statistical methods such as z-score normalization and Principal Component Analysis (PCA) to detect and filter out these anomalies. Removing outliers helped preserve the integrity of the dataset and allowed the model to make more accurate and stable predictions.

Most babies in the dataset were born weighing about **3200 grams**, but there were a few very low or very high weights, as low as **230g** and as high as **6300g**. These unusual values might be due to rare medical cases or possible mistakes in the data.

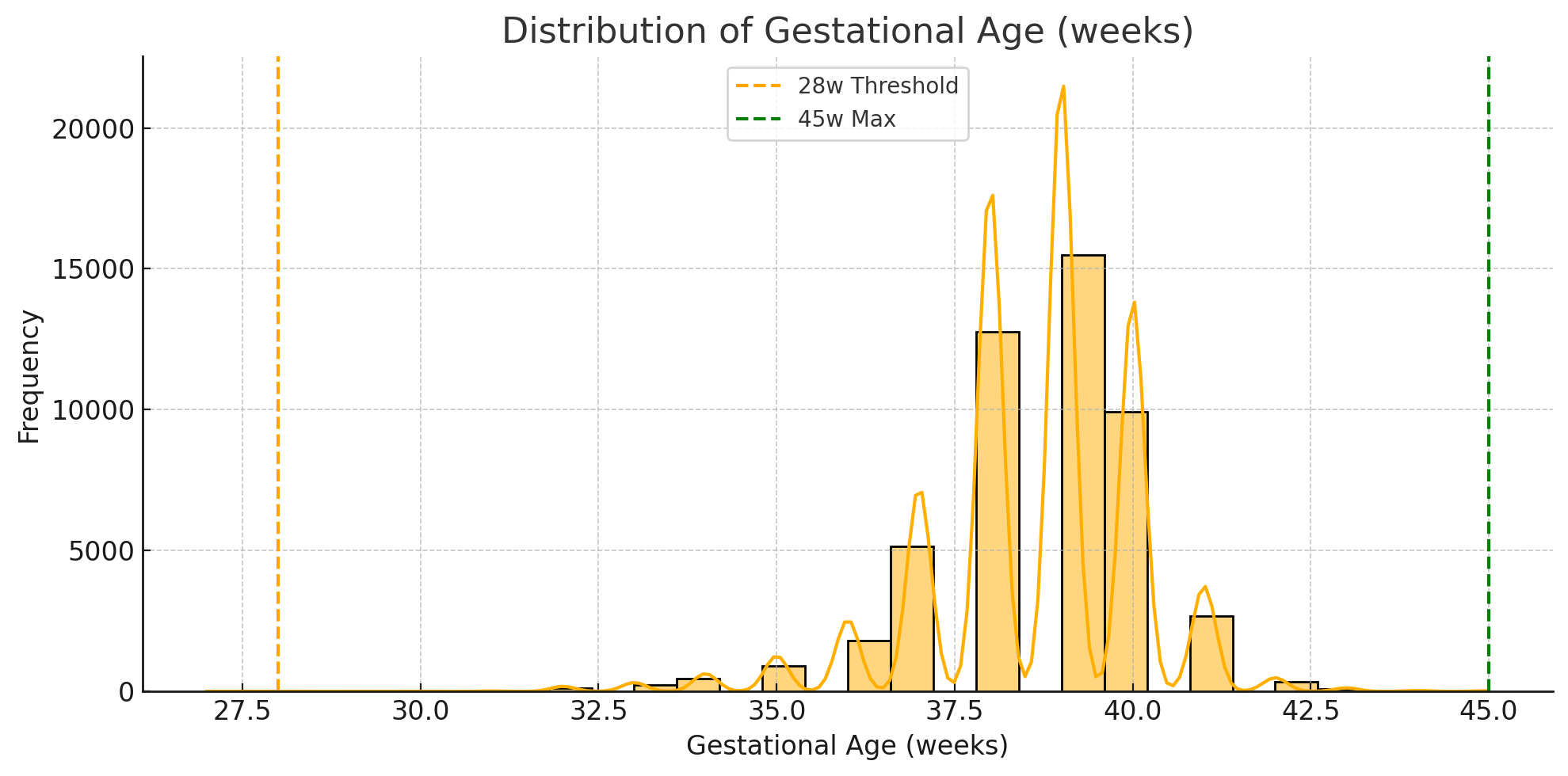
The average time of pregnancy was about **38.6 weeks**, which is considered full term. Some babies were born as early as **27 weeks**, which is very early and often comes with serious health risks.

We found that babies with very low or very high weights, or those born very early, had a higher chance of dying shortly after birth. This shows how important it is to watch these factors carefully in hospitals and clinics.



**Figure 3:***Distribution of Newborn Weight in the Dataset.*

The histogram shows a near-normal distribution centered around 3200g. Most weights fall between 2500–4000g. Red dashed lines mark the 500g and 1000g thresholds, highlighting medically significant low birth weights



**Figure 4:***Distribution of Gestational Age (weeks).*

The chart shows most births occur between 36 and 41 weeks, with peaks at full term (around 39–40 weeks). Dashed lines mark key thresholds: 25 weeks (viability limit) and 45 weeks (upper clinical boundary)

Most babies who had a low Apgar score at 1 minute improved quickly. About **97% of them reached a normal score by 5 minutes**, showing that many newborns bounce back shortly after birth. Only around **3% stayed in the low range** at 5 minutes. We figured this out by comparing the babies with low scores at both time points and seeing how many improved.

The number of previous pregnancies **num\_gestations** had a very high correlation (0.88) with the number of live births, which makes sense - women with more pregnancies often had more live births. Similarly, the number of normal deliveries **num\_normal\_labors** was strongly correlated (0.74) with the number of live births, reflecting that women with more vaginal births typically had more children overall. Lastly, we found a strong correlation (0.69) between the actual number of gestational weeks **num\_gestational\_weeks,** and the categorized pregnancy duration **tp\_pregnancy\_duration**, confirming that the categorical classification aligns well with the continuous week count.

**6.4 Model Selection and Training**

Throughout the model selection phase, we experimented with several algorithms. Logistic Regression was used as a baseline due to its simplicity and interpretability. Random Forest and XGBoost showed strong performance, particularly in handling non-linear relationships and imbalanced data. Neural Networks (MLP) were also tested and required more fine-tuning, including adjustments to layer size, dropout rate, and learning rate. In our trials, Random Forest achieved the highest recall without significant overfitting, while XGBoost performed best in terms of F1 score. These results helped us select the best model for clinical use while maintaining explainability.

**Logistic Regression** Logistic Regression is a statistical model used for binary classification. It estimates the probability that a given input belongs to a particular category using a sigmoid function. We used it as a baseline due to its simplicity and interpretability. It helped us validate that more complex models were truly needed and provided a reference point for performance. We chose Logistic Regression as a baseline model because it is simple, fast, and easy to interpret. In medical applications, especially early-stage research, it's important to start with a model that offers transparency. Although we knew it might not capture complex relationships in the data, it helped us establish a benchmark and ensured that more advanced models were truly adding value.

It is a simple and widely used model in healthcare research. It provides clear and interpretable results, which is important when explaining predictions to doctors and clinical staff.

**Random Forest** Random Forest is an ensemble learning method that builds multiple decision trees and combines their outputs to improve accuracy and stability. It is good at handling non-linear patterns and noisy data. We selected it because it works well with tabular healthcare data and is less prone to overfitting. It also helped us identify which clinical features were most important. Random Forest was selected due to its robustness and ability to handle non-linear relationships between features. It works well with tabular data like ours and is naturally suited for datasets with missing values or noise. Importantly, it performs internal feature selection and provides clear feature importance rankings. Given our imbalanced dataset (with few low Apgar score cases), we expected Random Forest to maintain good recall while avoiding overfitting.

It is a powerful tree-based model that works well with both numerical and categorical data. It is also good at handling missing values and imbalanced data, which is helpful for our dataset. In addition, it gives us a way to measure feature importance, so we can understand which factors are most related to Apgar scores and neonatal outcomes.

**XGBoost** XGBoost (Extreme Gradient Boosting) is a powerful algorithm that builds decision trees sequentially, where each new tree corrects errors made by the previous ones. It is optimized for speed and performance and is especially effective on structured data. We chose XGBoost because it performs well on imbalanced datasets and has been successful in similar medical prediction tasks.

XGBoost (Extreme Gradient Boosting) is a high-performance ensemble algorithm known for its excellent results in classification tasks. We included it because of its ability to deal with imbalanced data, manage missing values, and deliver high accuracy. Based on previous studies in the field, we anticipated that XGBoost would yield strong F1 scores, particularly in detecting rare but critical neonatal risks.

**XGBoost** is a more advanced tree-based model that often achieves higher accuracy than Random Forest. It uses boosting to correct errors from previous trees, which makes it very effective for complex patterns in the data. XGBoost is also fast and efficient for large datasets like ours.

**MLPClassifier (Neural Network)**

A Multi-Layer Perceptron (MLP) is a basic type of neural network composed of multiple fully connected layers. It is commonly used to model complex, non-linear relationships between input features and target outcomes. In medical prediction tasks, MLPs have the potential to uncover patterns that simpler models might miss, especially when trained on large datasets. While this model offers strong predictive capabilities, it often requires careful tuning-such as adjusting the number of layers, neurons, and dropout rates-and tends to be less interpretable than tree-based models.

Using MLP required careful setup, like choosing how many layers and neurons to include, and adding dropout to avoid overfitting. But it allowed us to explore whether a more advanced model could improve predictions. Neural networks are especially helpful when the connections between the features and the outcome aren’t simple or straight.

This type of model is often used to compare the capabilities of deep learning approaches with more traditional models such as logistic regression or decision trees.

**Neural Networks** can capture more complex patterns in the data. They are especially useful when the relationships between input variables and the outcome are not linear. Since our dataset is large, we had enough data to train a deep learning model and compare its performance to the simpler models.

By comparing these four models, we aimed to balance interpretability, performance, and real-world usability. This approach helped us find the best-performing model while still considering how practical it would be for doctors to understand and use.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Strengths | Weaknesses | Best Performance Metric | Contribution to Our Project |
| **Logistic Regression** | Simple and fast; Highly interpretable; Good baseline for comparison | Limited to linear relationships; Lower performance on complex data | Baseline Accuracy | Helps establish a benchmark and provides clear explanations for clinicians; ensures more complex models add value |
| |  | | --- | | **Random Forest** | | Handles non-linear relationships; Robust to outliers; Good feature importance explanation; Suitable for imbalanced data | Slower with large datasets; Less interpretable than logistic regression | High Recall | Handles imbalanced data effectively and highlights important features like gestational age and birth weight |
| **XGBoost** | Very high accuracy; Excellent with imbalanced and structured data; Fast and efficient; Handles missing values | Requires careful hyperparameter tuning; Lower interpretability | Highest F1 Score | Captures complex patterns and improves detection of rare low Apgar cases; effective with large medical datasets |
| **MLPClassifier** | Learns complex relationships; Suitable for non-linear problems; Can outperform simpler models if tuned well | Less transparent; Needs tuning (layers, dropout, etc.); Risk of overfitting | Moderate AUC-ROC and Recall | Explores deep relationships in the data; offers potential improvements over tree-based models |

**Table 1:** A comparative summary of the machine learning models evaluated in this project.

This table outlines the strengths, weaknesses, and best performance metrics for each model tested, as well as their specific contribution to the Apgar score prediction task. The comparison highlights how different models handle challenges such as imbalanced data, interpretability, and complexity of clinical patterns. This analysis helped guide the selection of the most suitable model for real-world medical application.

Among the various models we evaluated, the most promising ones so far are Random Forest, XGBoost, and the Keras Deep Neural Network (DNN). Each of these models demonstrated strong performance in handling the complexity and imbalance of our clinical dataset. While traditional models like Logistic Regression offered interpretability, their performance was limited, especially with non-linear relationships in the data.

The Keras DNN, as a deep learning model, stands out due to its ability to capture complex patterns between maternal and fetal features. However, it requires more careful tuning and is less interpretable for clinical use. On the other hand, XGBoost provided the highest F1 score, making it ideal for accurately detecting rare but critical cases like low Apgar scores. Random Forest also performed well, with high recall and solid feature importance visualization, making it a strong candidate for real-world implementation.

Overall, while AI-based models such as Keras DNN show great potential in terms of predictive power, tree-based models like XGBoost and Random Forest currently lead our selection due to their balance between accuracy, robustness, and explainability-especially when supported by SHAP value analysis.

In addition to evaluating the models based on technical metrics such as recall, F1 score, and AUC-ROC, we also considered their practical suitability for clinical use. While deep learning models like the Keras DNN showed strong predictive power, their complexity and limited interpretability could pose challenges in real-world medical settings. In contrast, models like Random Forest and XGBoost offered a better balance between performance and explainability. This makes them more appropriate for integration into clinical workflows, where transparency and trust in the model’s predictions are essential for adoption by healthcare professionals.

**6.5 Evaluation and Testing**

To make sure our model is both accurate and useful in real healthcare settings, we used a thorough validation process that focused on four main areas: data quality, model performance, practical usefulness, and technical reliability.

First, we looked at the dataset carefully. We checked for missing values, made sure the medical numbers made sense (like the mother’s age and number of pregnancies), and checked that related information didn’t conflict. Because neonatal death happened in only a small number of cases, we used a method called SMOTE to balance the data. This helped the model learn better from these rare but important cases.

To see how well the model works, we split the data using a method called a stratified train-test split. We also used k-fold cross-validation to make sure the results were stable. We checked how good the model was using accuracy, recall, precision, F1 score, and AUC-ROC. We also looked at confusion matrices to understand where the model made mistakes. To make sure the model's predictions made sense medically, we looked at which features were most important and used SHAP values to explain the results.

To make sure our model's predictions made sense in a real clinical setting, To interpret the model’s predictions and ensure clinical trust, we used SHAP (SHapley Additive exPlanations), an explainability technique based on cooperative game theory. SHAP assigns each feature an importance value that reflects its contribution to the final prediction, based on the concept of Shapley values from economics. This method considers all possible combinations of features and computes how much each one adds to the prediction.

In our case, SHAP allowed us to visualize and understand which features-such as birth weight, gestational age, and maternal age-had the most influence on predicting low Apgar scores or neonatal death. Compared to traditional feature importance methods, SHAP offers more consistent and locally accurate explanations.

We used SHAP plots to verify that the model was focusing on clinically relevant variables, which increased our confidence in the predictions and aligned with medical domain knowledge.

**Figure 5: Illustration of how SHAP explains a single prediction.**

This is a generic example to demonstrate the concept of SHAP values. It does not represent real data from our dataset but helps visualize how individual features (like age or blood pressure) contribute to a model’s output.<https://github.com/shap/shap>

**6.6 Key Predictive Insights**

To understand which features might be important for predicting Apgar scores and neonatal death, we used tools like correlation analysis, feature importance from machine learning models, and SHAP values. So far, we did not find any single feature with a clear, strong impact on Apgar scores. However, by using these models, we hope to uncover more subtle patterns and relationships between multiple features that together could help predict outcomes. This approach may reveal complex interactions that are not visible through simple analysis, helping us better understand which factors influence Apgar scores in real-world settings.

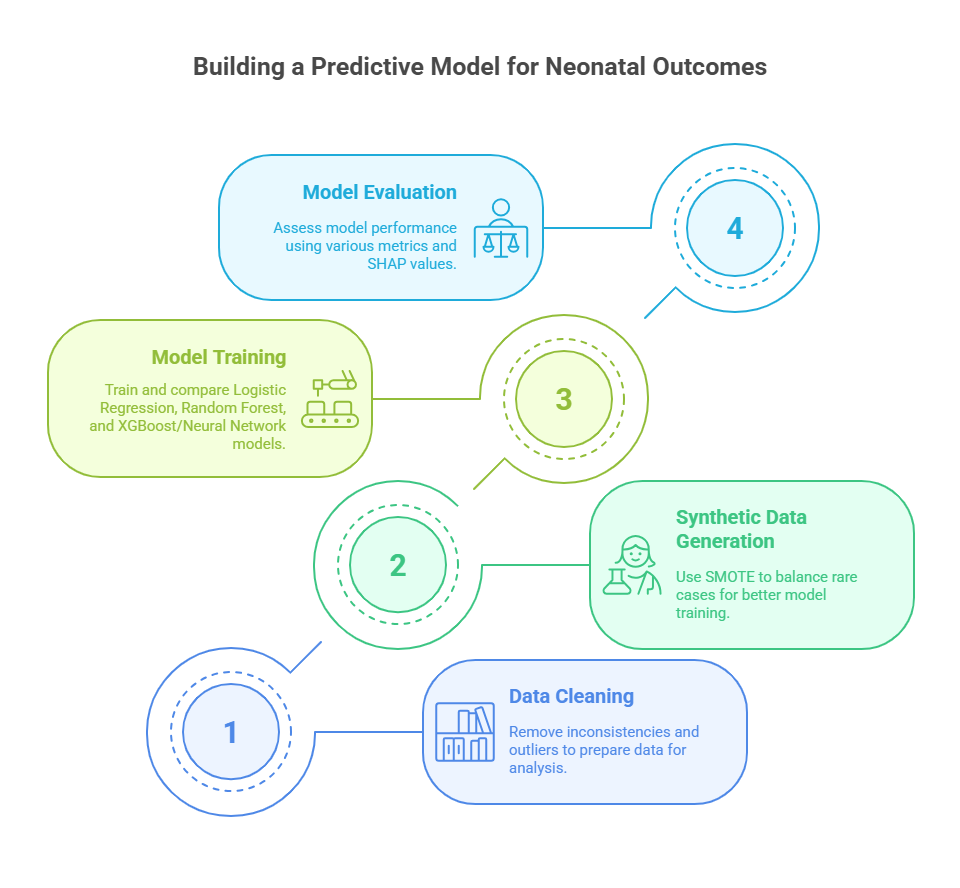
**6.7 Innovation and Contribution**

This project stands out due to its practical approach and clinical relevance. While many studies focus on retrospective analysis of neonatal outcomes, our work aims to create a predictive tool that can support real-time decision-making during labor or even in prenatal checkups. Our personal contribution includes designing a two-part training approach that allows the model to function even when some newborn data is missing-making it more robust for real-world scenarios. We also integrated advanced techniques such as SMOTE for balancing rare cases and SHAP for explainability, ensuring both fairness and interpretability. By combining machine learning with clinical insight, we propose not just a technical model, but a medically meaningful system that could be integrated into hospital workflows and improve perinatal care.

**real-time clinical application**.

One of the most innovative aspects of our project is the potential for real-time clinical application. Most existing Apgar score studies focus on retrospective analysis-looking at the baby's condition after birth. However, our project aims to shift this timeline forward by enabling pre-birth risk prediction using machine learning.

This means that during labor, or even before it, doctors could input maternal and pregnancy-related data into a simple interface and receive an instant prediction about the newborn’s expected Apgar score. By integrating the model into hospital systems or electronic health records (EHRs), it could become a real-time decision support tool, helping identify high-risk cases earlier and plan appropriate interventions. This forward-looking approach enhances both personalization and preparedness in perinatal care, making it a truly novel contribution to the field

**Figure 6:***Workflow for Building a Predictive Model for Neonatal Outcomes.*

The diagram outlines four main steps: data cleaning, synthetic data generation using SMOTE, training multiple machine learning models ( Logistic Regression, Random Forest, XGBoost, Neural Networks), and evaluating performance using metrics and SHAP values.

To build our predictive model for neonatal outcomes, we followed a structured, four-stage workflow. First, we performed data cleaning to remove inconsistencies, handle missing values, and eliminate outliers, ensuring that the data used for training was reliable and accurate. Next, in the synthetic data generation phase, we applied the SMOTE technique to balance the dataset by oversampling rare but clinically critical cases, such as low Apgar scores. This step was essential to prevent the model from ignoring minority outcomes. In the model training phase, we developed and compared multiple machine learning models, including Logistic Regression, Random Forest, XGBoost, and Deep Neural Networks, evaluating their ability to capture complex relationships in the data. Finally, during model evaluation, we assessed performance using metrics such as recall, F1 score, and AUC-ROC, and used SHAP values to interpret how different features contributed to the predictions. This complete pipeline helped us design a model that is not only accurate but also explainable and clinically relevant.

### **7. Future Work**

While this project builds a strong starting point for using machine learning to predict Apgar scores before birth, there’s still a lot we can improve. During our work, we found several ways future research could make the system even more accurate, useful, and ready for real hospital settings.

One important next step is to run all three models we tested and compare them to find which one works best and gives the most reliable results..

**7.1 Using More Data and Richer Information** Our current model was trained on a single dataset with basic medical features, like maternal age and fetal weight, and more. Going forward, we’d like to include data from more hospitals and possibly other countries to make sure the model works well in different settings. Adding other types of information-such as prenatal ultrasound results, maternal health history, or lifestyle factors-might help the model spot risks more clearly.

We plan to build a model that, based on information about the mother, will estimate the newborn's Apgar 1 and Apgar 5 scores, as well as the risk of neonatal death.

The model will be trained in two parts - one using records that include Apgar scores, and one without them - so it can learn to make predictions even when some information is missing.

The goal is to create a tool that gives doctors early insight into the baby’s health before birth, by using maternal data to predict the expected condition of the newborn.It will also help fill in missing newborn information based on the mother's details.

**7.2 Focusing on Specific Risks**

Right now, our model predicts whether a baby is likely to have a low Apgar score, but it doesn’t say *why*. Future work could look at predicting specific causes, like breathing problems or complications from a premature birth. If doctors know not just *that* something might go wrong, but *what* might go wrong, they can prepare even better.

**7.3** **Trying Smarter Algorithms**

We started with solid, well-known models like Random Forest and basic Neural Networks, but there are many newer techniques we haven’t explored yet. Deep learning and more complex ensemble models might help uncover patterns we’re currently missing. We’re also interested in using explainable AI, which could help doctors understand *why* the model gave a certain prediction, making it easier for them to trust and use the tool

7.4 **Making It Work in Real Time**

In the future, we’d love to integrate it directly with hospital systems, so predictions can be made instantly during labor. Imagine a doctor entering patient data into an electronic health record and immediately getting a clear picture of the baby’s expected condition. That kind of integration could make the model a real part of the clinical workflow

7.5 **Testing It in Real Hospitals**

One of the most important next steps is testing the model in real hospitals. It’s one thing to have good results in development, but another to see it perform under real-world conditions. We hope future studies can run clinical trials to evaluate how much the tool helps and if it improves outcomes for newborns.

7.6 **Thinking About Ethics and Privacy**

As this kind of technology gets closer to real use, we also have to think carefully about how it’s used. That includes protecting patient privacy, avoiding algorithmic bias, and making sure patients and providers understand how the tool works. These are not just technical issues-they’re human ones, and they’ll need thoughtful attention as the project moves forward.

7.7 **Future Integration with Clinical Systems**

Looking ahead, one of the most impactful directions for this project is the integration of our predictive model into clinical decision support systems (CDSS) or electronic health record (EHR) platforms. By embedding the model into the digital infrastructure already used in hospitals, healthcare professionals could receive real-time risk assessments during labor or prenatal checkups. For example, when a doctor inputs routine maternal and fetal data into the system, the model could immediately provide an Apgar score risk prediction-flagging high-risk cases and prompting early intervention. This kind of seamless integration could greatly enhance the usability, timeliness, and clinical value of the model, making it a practical tool rather than just a research outcome.

7.8 **Opportunities for Further Improvement**

While the current model shows promising results, there are several ways it can be further improved. First, expanding the dataset to include more diverse populations and larger sample sizes could enhance model robustness and generalizability. In addition, integrating image-based data-such as fetal ultrasound scans or CTG (cardiotocography) results-may enrich the feature space and allow the model to capture important visual patterns that are not represented in tabular data. Finally, collaborating with hospitals to collect real-time clinical feedback could help fine-tune the model for practical use, ensuring it aligns with clinical workflows and addresses real-world edge cases.

**References**

**[1] Apgar, V.** (1953). A proposal for a new method of evaluation of the newborn infant. *Current Researches in Anesthesia & Analgesia, 32*(4), 260–267.

**[2] Casey, B. M., McIntire, D. D., & Leveno, K. J.** (2001). The continuing value of the Apgar score for the assessment of newborn infants. *New England Journal of Medicine, 344*(7), 467–471.<https://doi.org/10.1056/NEJM200102153440701>

**[5] Lee, H., et al.** (2023). Predicting early mortality and severe intraventricular hemorrhage in preterm infants using machine learning models. *Scientific Reports, 13*(1), 1–10.<https://www.nature.com/articles/s41598-024-61749-1>

**[4] Nourani, M., et al.** (2022). Distributed big data analytics method for the early prediction of neonate Apgar score. *Journal of Big Data, 3*(2), 23.<https://www.mdpi.com/2673-2688/3/2/23>

**[6] Rocha, E. da S., et al.** (2022). On usage of artificial intelligence for predicting mortality during and post-pregnancy: A systematic review of literature. *BMC Medical Informatics and Decision Making, 22*, 334.<https://doi.org/10.1186/s12911-022-02082-3>

**[7] Tarimo, C. S., Bhuyan, S. S., Zhao, Y., et al,Mdoe, P. F., et al..** (2022). Prediction of low Apgar score at five minutes following labor induction intervention in vaginal deliveries: Machine learning approach for imbalanced data at a tertiary hospital in North Tanzania. *BMC Pregnancy and Childbirth, 22*, 275.<https://doi.org/10.1186/s12884-022-04534-0>

**AI Tools:**

ChatGPT (OpenAI)

<https://openai.com/chatgpt/>

Prompts:

“Compare Random Forest and XGBoost in predicting Apgar score and explain when to prefer each model.”

"Explain the SMOTE algorithm in simple English for a medical presentation."

“Generate key insights from this document”

Claude (Anthropic)

<https://www.anthropic.com/index/claude>

Prompts:

“List 5 possible ethical concerns when applying AI to neonatal outcome prediction.”

“What are the limitations of using SMOTE for imbalanced medical datasets?”