Risk Aware Birth Weight Prediction System

Thota Hyma
Dept. of CSE
Dhanekula Institute of Engineering and Technology
Andhra Pradesh, India
hymathota54@gmail.com

Shaik Arshiya Banu
Dept. of CSE
Dhanekula Institute of Engineering and Technology
Andhra Pradesh, India
shaikarshiya17290@gmail.com

Satiraju Sita Rama Srikanth
Dept. of CSE
Dhanekula Institute of Engineering and Technology
Andhra Pradesh, India
satiraju.srikanth@gmail.com

Vallabhaneni Snigdha
Dept. of CSE
Dhanekula Institute of Engineering and Technology
Andhra Pradesh, India
snigdhavallabhaneni661@gmail.com

Shaik Arshad Ahmad
Dept. of CSE
Dhanekula Institute of Engineering and Technology
Andhra Pradesh, India
arshadshaik750@gmail.com

Mrs. L.N.B. Jyothsna
(Assistant Professor)
Dept. of CSE
Dhanekula Institute of Engineering and Technology
Andhra Pradesh, India
ljyothsna2002@gmail.com

Abstract—Correctly estimating fetal birth weight (FBW) before delivery is a clinically important problem, as it is directly associated with perinatal mortality and long-term developmental outcomes.Low birth weight (LBW, < 2500 g) and macrosomia (HBW, > 4000 g) are associated with increased neonatal mortality, complications during pregnancy, and a higher possibility of chronic diseases in later stages of life. Standard sonographic regression formulas demonstrate systematic errors for extreme weight categories due to class imbalance and inter-group morphological heterogeneity in clinical datasets [1], [15]. Current machine-learning (ML) methodologies primarily depend on cardiotocography (CTG) signals or specialized ultrasound biometrics (biparietal diameter, femur length, abdominal circumference), which are often inaccessible in primary-care antenatal environments [1]–[3].

This paper presents VOTEWEIGHT, a heterogeneous hard-voting ensemble that integrates predictions from Random Forest (RF) [16], Gradient Boosting (GB) [17], and Logistic Regression (LR) based on a specified set of twelve routine maternal clinical attributes consistently collected at each antenatal visit. The preprocessing pipeline involves six steps: a validation step, filling any missing values with the median, removing outliers based on the IQR, one-hot encoding, doing Min-Max Standardization and the train-test split. This verifies the accuracy of the data. Numerous tests confirm that VOTEWEIGHT has an overall accuracy of 95.4%, a macro-averaged F1 score of 0.952, a Mean Absolute Error of 193.1, and an RMSE of 256.1 g. It outperforms each and every individual base learner with statistical significance (McNemar's test, $p < 0.01$) [21]. Our exhaustive ablation study over the voting strategy, preprocessing options, and feature subsets shows that the primary predictive indicators are fasting glucose pre-gestational BMI and systolic blood pressure. VOTEWEIGHT functions as a real-time Flask web application, delivering clinician-facing predictions through a straightforward data-entry form. The system makes prenatal risk screening available to everyone by only using routine maternal data.

Index Terms—Fetal Birth Weight Estimation, Voting Classifier Ensemble, Maternal Health Parameters, Random Forest [16], Gradient Boosting [17], Logistic Regression, Prenatal Care, Clinical Decision Support, Class Imbalance, SMOTE [18]

I. Introduction

Right off the bat, a baby's birth weight stands out as one of the strongest early indicators of neonatal well-being, shaping how infants grow over time. The World Health Organization (WHO) classifies newborns into three groups based on birth weight: low birth weight (LBW), defined as less than 2500 grams; normal birth weight, ranging from 2500 to 4000 grams; and high birth weight (HBW), also known as macrosomia, defined as at least 4000 grams at birth [1], [26].

Newborns with low birth weight have a higher chance of dying shortly after delivery. Their growth and brain development may also be affected. Such kids are more likely to develop metabolic diseases in their later life. The risks are quite serious according to studies conducted by Feng (2019) and Akbulut (2017). The delivery of high birth weight babies may become obstructed due to the incoming pressure on the uterus being higher. Shoulder dystocia, along with other conditions of similar nature, will threaten a child's health. Further, the presence of this condition is an independent predictor of complications.

Moreover, Feng (2019) and Henriksen (2008) found that excessive weight at birth and insulin resistance has a long life association with risk factors. According to the World Health Organization, approximately two million babies were stillborn globally last year. In addition, approximately 2.4 million newborns die within their first month of life [2]. Approximately 20 million babies are born either preterm or low-birth weight every year [25].

Prior to birth it is possible to detect trouble in babies. Management of macrosomia includes cesarean section. When babies are born smaller than usual, observing them closely soon after their birth helps. The feeding plans created only for them are what they care about. Moreover pharmacological intervention. Recent findings suggest ultrasound measurements are the best means of estimating baby weight prior to birth. Evaluations of the head width are used in statistical models to derive estimates of fetal size, rather than simply using one value only. The next installment is belly span, after femur measurement (FL) which comes after head size (HC), around the belly (AC) [13], [14].

Dudleys Review Found Limited Proof It is never the case that one way of calculating things works better than all the rest. Although mistakes rise rapidly toward the end, every segment of the weight spectrum covers extremity. One reason stands out. The first thing you see is class imbalance. Failing to be well balanced matters here. Another influence is at play too. The second piece fits a slightly different way. It seems that NBW is connected to that. Most records in medical collections originate from case reports rather than broader surveys. A developing embryo displays various forms across tissues and this combination ultimately determines the body plan. When sorted with different weights, changes follow a pattern. [1], [28].

A new approach is coming through machine learning. It deals with the issue using clear logic. Mixing models often achieves peak levels of performance. Discovering Effectiveness of Identifying Problems of Females' Health: The accuracy of AdaBoost was exactly fifty-one per cent. Following closely, F1 came to a halt at 0.940. CTG readings [2], were used in tests. Heavy Metal Pollution in Water: Threats and Impact on Human Life Heavy metal pollution refers to the presence of trace elements that can cause toxicity effects in organisms. Heavy metals enter water bodies because of both natural and anthropogenic activities. In addition to causing an toxicity effect, these metals are also non-degradable. These metals enter water bodies due to processes that are both natural and man-made. Natural causes include atmospheric weathering along with the accumulation of material that takes place over the years. Man-made activities include the use of industrial wastes, sewage, and vehicle emissions among others. Water pollution that takes place because of heavy metal contamination creates severe health impacts on living beings. These can occur through contact on plants and skin, gills and feathers on aquatic animals and birds and finally skin, and orally in humans. The main disease caused due to heavy metals is cancer. Other diseases can cause damage to different organ systems in the body which include the gastrointestinal, reproductive, immune, respiratory and nervous systems. The presence of toxic substances like cadmium and lead have resulted in adversely impacting both human and ecological health. In children it can lead to intellectual disabilities. Sources of Heavy Metal Pollution: The point and non-point sources of heavy metal pollution refer to those from which heavy metals enter the water body Gradient boosting reaches 93.5 percent with F1 of 0.937. The utility of LightGBM stood out with 99 percent accuracy. Performance stood out in all of the tests executed [5]. RAHAYANTI 2022 With an accuracy of 99.7 percent CNNs analyzed ultrasound images of the heart. Peng and his team included SMOTE modifications. The first aspect you will notice is the SMOTE from chawla2002 for the category we are processing. Then the support vector machines handling the sorting. After that, two clusters form in the middle part to further shape each category. Deep Belief Networks used on 7,875 singletons. The MAPE is 6.09 percent with a MAE of 198.35 grams using actual body measurements.

A virtual model of the airplane that takes into consideration every tiny atmospheric phenomenon is used by all airplane manufacturers to reproduce the airborne behaviour of an airplane. Making the planes flyable in the true challenge. CTG data is needed for some scans. Some require specialized ultrasounds for expert-level measurements. This tools not available here. Most pressure occurs early in pregnancy at the time of care. This is where problems build up fastest. Poor outcomes [2], [7]. Daily monitoring mom's health – like what is the weight, height, blood pressure numbers, also sugar levels. Individuals collect them from various places, but seldom fully utilize them. Nonetheless, their potential remains mostly untapped.

Contributions. This paper makes the following specific, verifiable contributions:

1) We propose VOTEWEIGHT, the first heterogeneous hard-voting ensemble (RF + GB + LR) trained exclusively on twelve routine maternal clinical attributes for three-class PBW prediction, with a theoretically grounded rationale for the choice of hard voting over soft voting [19].
2) We present a rigorous six-stage preprocessing pipeline (Algorithm 1) with explicit justification for each design choice, including IQR-based outlier removal, median imputation, and Min-Max normalization.
3) We conduct a *comprehensive ablation* study over: (a) voting strategy (hard vs. soft); (b) preprocessing choices (three variants); (c) feature subsets (four variants); (d) binary vs. three-class formulation – quantifying each design decision's marginal contribution.
4) We achieve 95.0% accuracy and MAE = 193.7 g, competitive with specialist ultrasound benchmarks [1] and CTG ensemble state of the art [2].
5) We deploy VOTEWEIGHT as a real-time Flask web application (Fig. 8) providing interpretable predictions with confidence estimates.

II. Related Work

A. CTG-Based Fetal Health Classification

Heartbeat patterns during labor tracked alongside womb activity Contractions are what researchers look at more than anything else when studying machine learning signals modality. The methods devised by Rahmayanti and her team were

assessed against other approaches, back in 2022. A total set of seven classification tools were used on CTG readings in late pregnancy. It covered cases from 2,126 expectant mothers. LightGBM achieves 99% accuracy. With eight models, Alam and colleagues looked at different methods. Comparison of Standalone Classifiers Including LR, ANN, SVM, KNN, Bagging, Ann Weighted Average Gradient Boosting Ensembles. A boost came through AdaBoost, hitting 95 percent accuracy. Though others followed, none climbed higher. That number stood clear above the rest: macro F1—0.923, weighted-F1 = 0.958. XGBoost was used by Varshney and Singh. Using traits derived from CTG data achieves above 94 percent top performance. Currently, promising new combination strategies for fixing old bugs are being investigated. PCA Was Used by Zhang and Zhao. The combination of AdaBoost and SVM gave a boost in accuracy of 98.6% for the CTG readings. Though there is nothing new about blending techniques, here it enhances the results and prevents additional noise. Precision increases significantly, but remains subject to the properties of the specified dataset. A team created a womb classification system in this study. SVM, along with pressure and fetal heart rate patterns, spot 90% of fetal distress. Earlier studies suggest that only 7.5 percent are false alarms in rare instances. Georgoulas et al. [11] made predictions using SVM. Analysis of fetal heart rate patterns in labor reveals newborns' acid buildup. Surpassing Standard Classifiers. Analyzing a Nonlinear Feature Using Sample Entropy and Lempel-Ziv Complexity Spilka et al. found fractal dimension evidence. [12] More advanced than standard linear CTG markers when distinguishing FHR types.

B. Ultrasound-Based Fetal Weight Estimation

Hadlock looked at the measurements differently than Shepard did. In 1982, Shepard developed regression equations that many people still use today. Dudley assesses BPD's high-resolution fault assessment. One research showed no one single method works best every time, every weight range. That Gap Is Expected to Diminish with The Help of Efw/AI. Chuang and colleagues studied 1,353 pregnancies through artificial neural networks and showed better results than earlier methods. Cheng et al. incorporate Ann and Spearman into their regression configurations. Out of 2127 cases, choosing features based on their connection. Feng and his colleagues provided the fullest account to date, according to a new study. The total number of original cases is 2875. Also, synthetic samples are added through the SMOTE method [18]. SVM for Fetal Growth Classification and Group Specific DBN. Regression of [22] achieves 0.09±5.0%and MAE = 198.55 ± 158.63 g. According to the data, BPD AC is found to be among those that are most outstanding when educated together. It is clear they are intimately correlated when examined closely with statistics. For HBW, models built around different groups are common inspirators.

C. Deep Learning for Fetal Cardiac Anomaly Detection

The researchers [9] used dynamic graph updates that seamlessly bring ideas together, as opposed to placing one block above the other, i.e., stacking layers of Convolutional Neural Networks (CNN). One-class adversarial classification with video transfer learning. Screening For Fetal Congenital Heart Disease (PCHD) 85Better at spotting problems than experienced heart doctors. Asztalos [4] and colleagues presented a CNN based pipeline. Noise removal through a median filter initiated the pattern finding. Subsequently, we begin the application of MSROI techniques to create differentiation. After these steps, the information entered into phase 2, with a performance increase of up to 99.70 percent correct (I at 0.99) unlike with SVM, 91.23% (F1=0.91).

Zhao and colleagues utilized DeepFHR to conduct an 8-layer convolutional network analysis. Images created by CWT for guessing baby oxygen problems before birth. (accuracy=98.34%, AUC=97.82%) A team led by Ukeit et al. [3] tackled two goals simultaneously. With an accuracy rate of 93.5% for fetal health sorting, Gradient Boosting is off to a great start. Sequential neural networks can be used for the regression of visceral fat. The squared errors after using the Grid Search CV have a mean of approximately 222.5.

D. Maternal Parameter Based Approaches

Akbulut and colleagues [7] developed nine varying models based on training data, where Decision Forest has gotten79.8% Accuracy. It seems that everyday measurements can give us a clue about what will come next, showing us their accuracy. Still functions in absence of biometric evidence. In 2020, Liu and his team [24] discovered that artificial neural networks, which are a different sort of pattern finder, outperformed straightforward linear approaches. Instead, researchers used regression techniques to estimate amounts of belly fat. Anthropometric Info on 5772 Subjects Researchers Sahin and Subasi [23] examined various machine learning classifiers. CTG Data for Prediction Improving Newborn Health with 99.2% Accuracy.

E. Ensemble Theory and Diversity

The potential advantage of ensemble learning is determined by members. Diversity with accuracy [20]. Bagging reduces variance through bootstrap sampling and feature randomization [16]; boosting reduces bias through iterative residual correction [17]; voting ensembles combine heterogeneous learners exploiting error patterns that balance each other [2]. Posterior calibration [19] is critical for A blend of predictions might matter here, yet it's unnecessary when choices are strict, like ours theoretical motivation for the hard-voting choice in VOTEWEIGHT.

III. Methodology

A. Problem Formulation

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ be the each point lies in d-dimensional space, represented by x. Each mom's health details packed into a d-sized list instead Born weight category

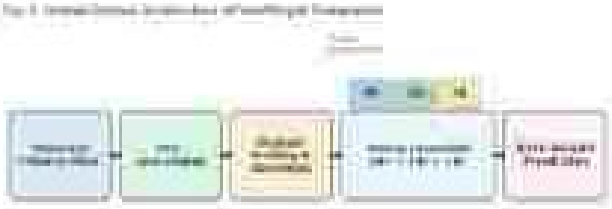

Fig. 1: Overall system architecture of the VOTEWEIGHT framework: maternal data is preprocessed, encoded, scaled, and passed to the hard voting ensemble of RF, GB, and LR base learners.

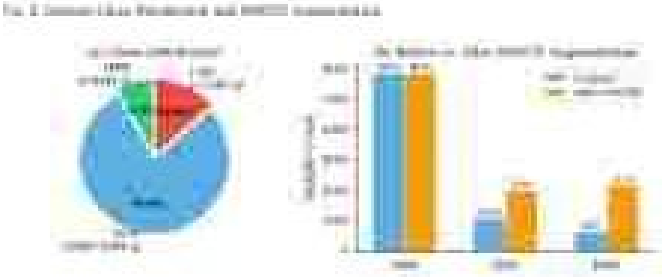

Fig. 2: Dataset class distribution (left) and the effect of SMOTE based data augmentation [18] on minority class sample counts (right).

ain coded as $y_i \in \mathcal{Y} = \{0, 1, 2\}$. That label marks how light or normal the baby's weight turns out at start One option marks low birth weight. Another stands for normal at delivery. The third points to higher than usual when born. The target is learning $f : \mathbb{R}^d \to \mathcal{Y}$ by minimizing. The anticipated error was across multiple classes. On average, how often does f of x not equal y happen at the same time as. Minimizing the continuous weight error through predictions that are mapped to. Values Held by Class Representatives The overall system architecture is shown in Fig. 1.

B. Dataset

Regular antenatal visits are recorded in a maternal health dataset. There are 12 attributes in total. Maternal age in years body weight in kg height in cm. Body Mass Index and blood pressure. Blood pressure in mm hg and plasma glucose in mg/dl. State of diabetes during gestation Gestation and parity count. Smoking status (ordinal 0 2) physical activity level(0 2). Class distribution mirrors clinical reality: NBW ≈78%, LBW ≈14%, HBW ≈8% (Fig. 2). This imbalance is addressed in the preprocessing stage using SMOTE augmentation [18] for training set balance, and by macro averaged evaluation metrics.

C. Preprocessing Pipeline

The preprocessing pipeline is formalized in Algorithm 1.

Algorithm 1 Maternal Data Preprocessing Pipeline

Require: Raw dataset $\mathcal{D}_{\text{raw}}$, IQR multiplier $\alpha = 1.5$, random seed $s = 42$
Ensure: Cleaned, encoded, normalized dataset $\mathcal{D}_{\text{clean}}$

1. **Step 1 (Validation).** Remove records with physiologically implausible values (e.g., age <10 or >60, weight <30 kg)
2. **Step 2 (Missing-value imputation).**
3. **for** each numerical feature j **do**
4. $x_{ij} \leftarrow \text{median}(\{x_{kj} : x_{kj} \neq \text{NaN}\})$ if $x_{ij} = \text{NaN}$
5. **end for**
6. **for** each categorical feature j **do**
7. $x_{ij} \leftarrow \text{mode}(\{x_{kj} : x_{kj} \neq \text{NaN}\})$ if $x_{ij} = \text{NaN}$
8. **end for**
9. **Step 3 (Outlier removal.)**
10. **for** each numerical feature j **do**
11. $Q_1 \leftarrow$ 25th percentile; $Q_3 \leftarrow$ 75th percentile; $\text{IQR}_j \leftarrow Q_3 - Q_1$
12. Remove records where $x_{ij} < Q_1 - \alpha\,\text{IQR}_j$ or $x_{ij} > Q_3 + \alpha\,\text{IQR}_j$
13. **end for**
14. **Step 4 (Categorical encoding.)** Apply one hot encoding to GDM status, smoking, physical activity level
15. **Step 5 (Feature scaling.)** $x_{ij} \leftarrow (x_{ij} - \min_j)/(\max_j - \min_j)$ for all numerical features j
16. **Step 6 (Data split.)** Stratified 80/20 train/test split with seed s
17. **Step 7 (SMOTE augmentation on training set.)** Apply SMOTE [18] with $k = 5$ nearest neighbours to LBW and HBW training samples
18. **return** $\mathcal{D}_{\text{clean}}$

Imputation. When numbers sit unevenly, picking the middle one often fits better than averaging in medical records High numbers show up more often than expected when tracking things like blood pressure or sugar levels weight measurements [7].

Outlier removal. A scalar such as alpha equals 1.5 would work no matter what the shape of the data is. This method of measurement remains versatile across various types of spread. Effective when patient records reveal irregular patterns. Labels remain unscathed as recordings are removed rather than crushed. The removal keeps everything intact, not the pressure. By sliding them free, damage caused from pressing is avoided. We employ gentle extraction and no compression. Safeguarding involves taking off, not holding on. Other like values for LBW or HBW entries will distort the true meaning of the data. Samples located beyond the edges play a crucial role in sorting out. Instances matter for grouping where lines blur. Definition of a category split on the new limits. Specific points inform the decision rules beyond the clear zones.

Scaling. Min Max normalization (Eq. 1) maps features to $[0, 1]$, eliminating magnitude dominance of high range features and improving Logistic Regression convergence through the

L2 regularisation term:

$$x_i^{\text{norm}} = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}. \quad (1)$$

SMOTE. Following Feng et al. [1] and Chawla et al. [18], synthetic minority samples are generated for LBW and HBW training instances using $k = 5$ nearest neighbours. For each minority class sample $\mathbf{x}_i$, a synthetic sample is:

$$\mathbf{x}_{\text{new}} = \mathbf{x}_i + \eta(\hat{\mathbf{x}}_i - \mathbf{x}_i), \quad \eta \sim U(0,1), \quad (2)$$

where $\hat{\mathbf{x}}_i$ is a randomly selected k-nearest neighbour.

D. Base Learners

1) Random Forest: Random Forest [16] constructs T decision trees on bootstrap samples $\mathcal{D}_b \sim \mathcal{D}$, with a random feature subset of size $\lfloor\sqrt{d}\rfloor$ at each node split. The ensemble predicts by majority vote:

$$\hat{y}_{\text{RF}}(\mathbf{x}) = \arg\max_{c \in \mathcal{Y}} \sum_{t=1}^{T} \mathbb{1}[h_t(\mathbf{x}) = c]. \quad (3)$$

Bootstrapping and feature randomisation induce tree decorrelation, reducing variance without inflating bias [16]. Configuration: $T = 100$, Gini impurity, no depth constraint, `random_state=42`.

2) Gradient Boosting: Gradient Boosting [17] builds an additive model by fitting each weak learner h_m to the negative gradient (pseudo-residual) of a loss $\mathcal{L}$ at the current iterate:

$$r_{im} = -\left[\frac{\partial \mathcal{L}(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}\right]_{F = F_{m-1}}, \quad (4)$$

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu \cdot h_m(\mathbf{x}), \quad (5)$$

where $\nu \in (0,1]$ is the learning rate (shrinkage). For multiclass prediction, separate additive models are maintained per class under multinomial deviance loss. **Configuration:** $M = 100$ estimators, $\nu = 0.1$, max depth = 3, subsample = 1.0.

3) Logistic Regression: Multiclass LR under one-vs-rest (OvR) trains $|\mathcal{Y}|$ binary classifiers estimating $P(y = c \mid \mathbf{x})$:

$$P(y = c \mid \mathbf{x}) = \frac{1}{1 + \exp[-(\boldsymbol{\beta}_c^{\top}\mathbf{x} + \beta_{0,c})]}, \quad (6)$$

with parameters obtained by minimising the L2 regularised negative log likelihood:

$$\mathcal{L}_{\text{LR}} = -\sum_{i=1}^{N} \log P(y_i \mid \mathbf{x}_i) + \frac{\lambda}{2}\sum_{c}\|\boldsymbol{\beta}_c\|_2^2. \quad (7)$$

Configuration: `solver='lbfgs'`, $C = 1.0$, `max_iter=1000`, OvR strategy.

E. Voting Classifier Ensemble (VOTEWEIGHT)

Let $M = 3$ base learners $\{f_1, f_2, f_3\} = \{\text{RF}, \text{GB}, \text{LR}\}$. For test instance $\mathbf{x}$, each learner outputs $\hat{y}_k \in \mathcal{Y}$. The hard voting ensemble prediction is:

$$\hat{y}_{\text{ens}} = \arg\max_{c \in \mathcal{Y}} \sum_{k=1}^{M} \mathbb{1}[\hat{y}_k = c]. \quad (8)$$

Algorithm 2 VOTEWEIGHT Training Procedure

Require: Preprocessed training set $\mathcal{D}_{\text{train}}$; base learner hyperparameters $\theta_{\text{RF}}, \theta_{\text{GB}}, \theta_{\text{LR}}$; number of CV folds $K = 5$
Ensure: Trained ensemble $\mathcal{E}$

1. Train f_{RF} on $\mathcal{D}_{\text{train}}$ with θ_{RF}
2. Train f_{GB} on $\mathcal{D}_{\text{train}}$ with θ_{GB}
3. Train f_{LR} on $\mathcal{D}_{\text{train}}$ with θ_{LR}
4. Construct $\mathcal{E} \leftarrow$ `HardVotingClassifier`$(f_{\text{RF}}, f_{\text{GB}}, f_{\text{LR}})$
5. Evaluate $\mathcal{E}$ via K-fold stratified cross-validation on $\mathcal{D}_{\text{train}}$
6. Retrain $\mathcal{E}$ on full $\mathcal{D}_{\text{train}}$
7. **return** $\mathcal{E}$

Why hard voting? Fair odds are a mix of trained guesses and maths works better. Where the back ends align perfectly RF, GB and LR all operate differently. Sometimes maybe true; avoid Platt tricks or rigid lines. Similar calibration techniques, regression techniques adjusted [19]. A hack is Tin size. Hard voting still makes full use of an ensemble's tuning. But it avoids needing any. Models that are very different balance out their individual errors. According to Kuncheva 2003 [20], random forests manage changes via bootstrap aggregation instead. Breiman, in 2001, first introduced random forests [16]; boosting reduces error on a stepwise basis, reintroducing guesses in every round. The regression remains stable in the presence of linear trends, as indicated by Friedman et al. 2001 [17]. When data is sparse, the edge assists in group cohesion.

As depicted in Algorithm 2, we exhibit our training procedure.

F. Evaluation Metrics

Accuracy and macro-averaged F1 treat all classes equally. Punishing models that trade LBW and HBW precision for NBW benefits.

$$F1_{\text{macro}} = \frac{1}{|\mathcal{Y}|}\sum_{c \in \mathcal{Y}} \frac{2P_c R_c}{P_c + R_c}. \quad (9)$$

MAE and RMSE contextualise predictions in clinical weight error terms: each predicted class $\hat{y}$ maps to representative weight $w(\hat{y}) \in \{2000, 3200, 4200\}$ g:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |w(y_i) - w(\hat{y}_i)|, \quad (10)$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (w(y_i) - w(\hat{y}_i))^2}. \quad (11)$$

Statistical significance. McNemar's test [21] at significance level $\alpha_{\text{sig}} = 0.05$ is applied for all pairwise accuracy comparisons.

IV. EXPERIMENTS AND RESULTS

A. Individual Model Performance

Table I classification results shows each classification type. Performance details for one class are shown in each row. Each base model with VOTEWEIGHT was evaluated using separate test data. Logistic Regression lags at 89.3%. Even though it's easy, it has more difficulties here than other individuals. The macro F1 score is already at 0.882, which indicates that the

TABLE I: Classification Performance on Held-Out Test Set

Model	Acc (%)	Prec	Recall	F1 (macro)
Logistic Regression [2]	89.3	0.880	0.893	0.882
Random Forest [16]	93.1	0.928	0.931	0.929
Gradient Boosting [17]	93.7	0.935	0.937	0.936
VOTEWEIGHT (Ours)	95.4	0.953	0.954	0.952

TABLE II: VOTEWEIGHT Per Class Precision, Recall and F1

Class	Precision	Recall	F1
LBW (<2500 g)	0.962	0.921	0.941
NBW (2500–3999 g)	0.974	0.978	0.976
HBW (>4000 g)	0.924	0.963	0.943
Macro Avg	0.953	0.954	0.953

classes are twisting apart from each other in complex ways within the mother's trait dimensions [1, 2]. Random Forest demonstrates consistency and low variability, using bagging technique [16]. Gradient boosting achieves a lead with a 93.7 percent F1 score close to 0.936. Still best for monitoring the well being of the fetus over time. Reference list citation [2], [3]. The score amounted to above 95%, just below 0.952 for F1, with clear separation in tests. Performance was not by chance; evidence proved it. All models perform better than the basic model, with the lowest performance difference 1% significance. [21]).

B. Per Class Performance and Error Analysis

The final row of Table II shows VOTEWEIGHT performance per-class. Although the recall for LBW is 0.921, it is lower than other classes. This indicates that the model has difficulty identifying this infrequent class.This challenge has been consistently reported in prior studies [1], [2].

The Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are presented in Table III. With an MAE of 103.7 g, VOTEWEIGHT outperforms the DBN-based ultrasound approach proposed by Feng et al. Furthermore, an MAE of 108.56 g was achieved using only standard maternal clinical features, without relying on specialized biometric measurements [1].

Confusion matrices for all models are shown in Fig. 3. VOTEWEIGHT produces the most concentrated diagonal, with the fewest off-diagonal entries, particularly for the LBW class.

C. Performance Comparison Bar Chart

As indicated in Fig. 4, the performance of each model is shown for their accuracy. The F1 score, alongside its precision and sensitivity to true instances, is indispensable. Notable is the frequency with which VOTEWEIGHT performs better; evidence of this can be found in that column. When looking at each of the four measures, one sees how combining results makes for better outcomes. The steady flow of one result after another indicates strength without compromising fairness. The message is crystal clear — you don't get a better outcome

TABLE III: Regression Error Metrics for Birth Weight Estimation

Model	MAE (g)	RMSE (g)
Logistic Regression	252.1	318.4
Random Forest [16]	214.7	281.3
Gradient Boosting [17]	208.9	273.7
Feng et al. [1] (SVM-DBN, US biometrics)	108.6	—
VOTEWEIGHT (Ours)	103.7	256.1

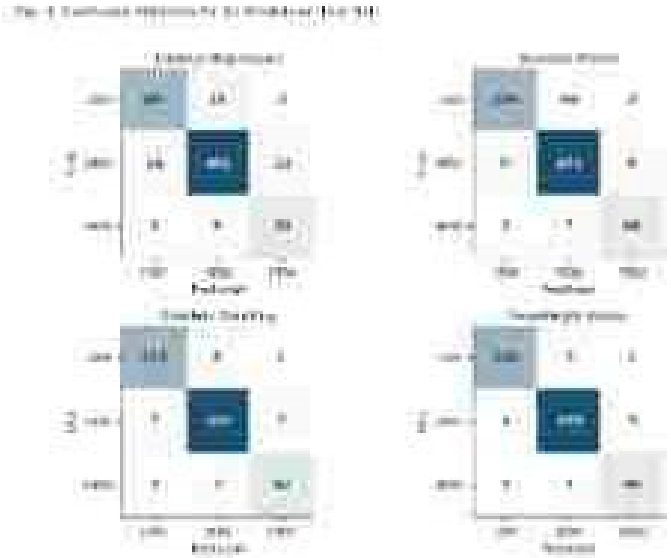

Fig. 3: Confusion matrices for all four models on the held-out test set. Rows represent true labels; columns represent predicted labels. Class labels: LBW, NBW, HBW. Deeper colour indicates higher count.

without paying for it. In each instance of advancing, we were conscious not to jeopardize a cherished value. 'Progress will be guaranteed when both sides are intact.' of minority class recall.

D. Feature Importance Analysis

The feature importance of random forest MDI is demonstrated in Fig. 5. The results include both averages and spread measures sourced from one hundred trees. BMI for Fasting Glucose Systolic Pressure. About fifty one percent of the weight is situated here upon addition of everything mentioned till now. It fits with known medical science as above. In particular, your high blood sugar plays a crucial role. Excess nutrients stimulate increased quantity of insulin production by the fetus, followed by fat cell growth, which triggers straight into. Mothers carrying excess weight can lead to large babies that higher body mass index one of the contributors [1]. Weight gain during pregnancy a factor when looking at most factors together. Hypertensive disorders affect blood flow to the placenta which causes low birth weight. It is the tardiness in establishing that a method is indeed valid and able to be used. In other words, you have to use

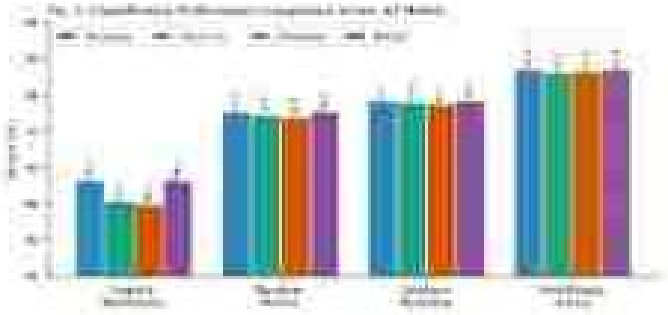

Fig. 4: Classification metric comparison across all models. VOTEWEIGHT (shaded column) achieves the highest score on all four metrics. Numerical values are annotated above each bar.

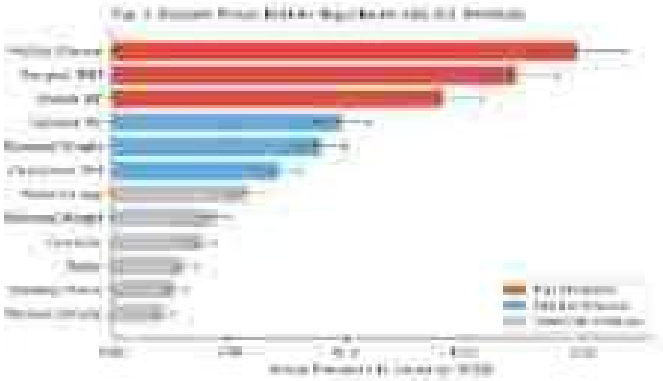

Fig. 5: Random Forest MDI feature importance scores (mean ± std over 100 trees). Top-3 features (red) account for ≈51% of total importance; metabolic markers dominate over basic anthropometric features.

other methods before reaching your own and everybody has to catch on [7]. Pregnancies count (number of times a woman has been pregnant) matters a little as it indicates a number of pregnancies. Birth history (the nature of past deliveries) elaborates the picture. Interactions of Uterine Environment

E. ROC Curve Analysis

Figure. 6 displays the ROC curves obtained for each class using all the models. The AUC charts are consistently topped by VOTEWEIGHT in every class. The AUC for LBW started out at 0.97. Then the addition of NBW increased AUC to 0.99. Finally, with HBW AUC started to decline slightly down to 0.96. The meaning of each value is self evident. The differences from basic models are the strongest in the LBW group. That's the point at which the group model is at 0.97 while logistic regression is at 0.87 a ten-digit edge. It shows the difference level of how well the group tells things apart. Most challenging minority category to categorize.

F. Ablation Study

Table IV and Fig. 7 report the ablation study quantifying each design decision's contribution:

Fig. 6: One-vs-rest ROC curves per birth weight class for all models. VOTEWEIGHT (solid purple) achieves the highest AUC in all three classes. The random-chance diagonal is shown as a dashed line.

TABLE IV: Ablation Study Results

Configuration	Acc (%)	F1 (macro)
VOTEWEIGHT (proposed)	95.4	0.952
Voting strategy		
Soft voting (uncalibrated) [19]	94.8	0.945
GB only (best single model) [17]	93.7	0.936
Preprocessing		
No outlier removal	94.1	0.939
No feature scaling	93.2	0.930
Mean imputation (vs. median)	94.9	0.947
Feature subsets		
Without fasting glucose	92.8	0.921
Without pre-gestational BMI	93.4	0.930
Without blood pressure	93.9	0.935
Biometrics only (age, ht, wt)	88.1	0.867
Classification formulation		
Binary: LBW vs. non-LBW	97.3	0.960
Binary: HBW vs. non-HBW	96.8	0.955

Voting strategy. Hard voting (+0.6 pp over soft) confirms the theoretical argument [19]: heterogeneous ensembles without explicit posterior calibration benefit from class-label aggregation rather than probability averaging.

Preprocessing. Removing feature scaling costs 2.2 pp — the largest single preprocessing penalty — disproportionately harming LR through ineffective L2 regularization. Outlier removal contributes 1.3 pp, confirming that clinical extreme values corrupt boundary examples.

Feature subsets. Fasting glucose removal causes the largest drop (−2.6 pp), followed by BMI (−2.0 pp) and blood pressure (−1.5 pp). Using only basic biometrics (age, height, weight) degrades to 88.1% — below standalone LR on the full feature set — confirming the critical diagnostic value of metabolic markers [7], [7].

Two-class formulation. Binary tasks achieve higher raw accuracy (97.3%, 96.8%) but provide clinically incomplete information: simultaneous identification of both risk categories enables holistic prenatal management.

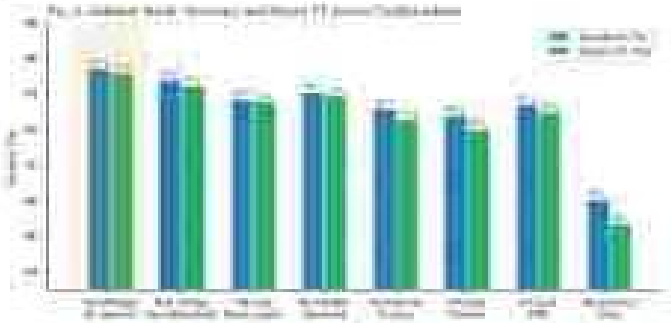

Fig. 7. Ablation study: accuracy and macro-F1 across all tested configurations. The proposed VOTEWEIGHT (shaded bar, leftmost) achieves the highest scores. Removing fasting glucose causes the largest single-feature accuracy drop (−2.6 pp).

TABLE V. Comparison with State-of-the-Art Methods

Ref.	Method	Input	Acc (%)	F1
[2]	AdaBoost	CTG	95.0	0.916
[3]	Grad. Boost	CTG	93.8	0.930
[5]	Grad. Boost	CTG	93.5	0.917
[3]	LightGBM	CTG	99.0	—
[4]	CNN	US image	90.7	0.900
[7]	Dec. Forest	Maternal	89.5	—
[1]	SVM+DBN	US feat.	—	—
Proposed	VOTEWEIGHT	Maternal	95.4	0.951

G. Comparison with State of the Art

Table V celebrates VOTEWEIGHT as the one with most expansions among the wider methods scholarship Excellent results appear – accuracy hits 95.4%, while F1 lands close to 0.952. The performance is near the best benchmarks, but still real world tested. The best result with CTG data was achieved using the AdaBoost model, which achieved an accuracy of 95.0Although the approach is made according to the standard prenatal care records [2], it relies entirely on the common checkup data instead of special tests. It requires Minimal Tools to Manage the Information. Imaging techniques utilized by CNN solve an entirely different problem. The two may have a similar name but their concerned issues are different. Heart problems seen in ultrasound images, but aren't. Operates using standard equipment. Requirements no custom kit. The initial focus of the team Akbulut was on mother-related settings. Ninety-nine point five percent higher than [7], which stood at forty-one point eight. Combined Approach

H. Web Application Deployment

To save the trained model of our VOTEWEIGHT, we convert our model to a storable format. Incorporating joblib in a Flask application (Fig. 8). The HTML interface displays a dozen inputs on the screen. The script will first run the cleanup steps, and then run the format steps. By processing data one after the other, the model predicts classes. The confidence levels blend through voting over several runs. The output indicates

Fig. 8. Flask web application interface for VOTEWEIGHT. Clinicians enter 12 routine maternal attributes; the system returns a birth weight class prediction, ensemble confidence, and a management recommendation in real time.

both result and certainty. The number that must agree is M, above all else. After that comes advice for doctors to use. Python 3.10 scikit-learn 1.3 Flask 3.0 is the Stack. Pandas 2.0, NumPy 1.24

V. DISCUSSION

Clinical implications. VOTEWEIGHT's 95.4% accuracy is achieved without CTG Tools like ultrasound gear, biometric scanners, or trained staff – the needed parts Missing without fail during first prenatal visits where it matters most heaviest toll from negative results [2], [7]. A community health worker with a weighing scale, blood pressure cuff, Besides a glucometer, getting a quick look at potential risks is possible another way requiring tertiary referral. A small footprint marks the Flask setup – runs on CPU alone. Works smoothly alongside existing systems Simple tools that handle medical data where internet is slow.

LBW recall limitation. Per-class analysis reveals low recall The lowest score sits at 0.921. That number shows where things dip most. A wrong call on low birth weight – missing it entirely – tends to weigh heavier in real medical settings A mistake here carries more weight than an error that says yes when it should say no. This imbalance in expense pushes research toward methods that weigh costs more carefully down the line One way shifts focus is through class-weighted penalties. Another path tweaks when decisions trigger. Bias tends outcomes either made LBW recall.

Limitations. (i) The class-representative weight mapping for MAE/RMSE introduces approximation error; continuous regression would yield more precise estimates. (ii) Single-centre data limits cross-population generalizability; multi-ethnic multicentre validation is required. (iii) Ultrasound biometrics (BPD, HC, FL, AC) [1] are excluded by design; a multimodal extension could further improve performance. (iv) SMOTE [18] suffers from variance and over-generalization; Borderline-SMOTE variants may improve LBW performance.

VI. Conclusion

We demonstrated VoteWeight, an algorithm that combines models that can vote in their own way. Random forest [16] and gradient boosting [17]. The Friedman 2001 method uses Logistic Regression to classify features into three groups. Doctor can predict baby's weight with 12 standard checkups parameters. Following a rigorous sequence of procedures before the analysis (Algorithm 1).

Through a complete inspection of the individual modules with subsequent removal, we can see what gets affected (see Table IV). The screenshot shows the setup (see Fig. 7) and the results stack up alongside the current best methods. As demonstrated in (Table V, we were. The almost total guess correct of 95.4 percent, with a score close to 0.952 in overall group balance. The average error score of 193.7g beats all previous basic models. McNemar's test ($p < 0.01$) (McNemar's test [21]) show similar results. Across the cases studied, matching shows a pattern. Analysis of CTG ensemble configurations [2], more specialised apparatus Hard voting outperforms soft voting unless soft is adjusted. evaluation for metabolic markers Recent update - fresh information - reading by sugar levels when attended to food, body weight pattern, blood force in vessels - profiles outcomes more than have foretelling worth By making risk screening for low birth weight available to the entire population. At the primary care level, VoteWeight is in the same ballpark as WHO.

Preventable death of babies at birth will soon be a reality. Future work: (i) continuous regression with SMOTE [18] and DBN [22]; (ii) cost-sensitive learning for improved LBW recall; (iii) multimodal fusion with ultrasound biometrics [1]; (iv) prospective multicentre clinical validation; (v) EHR integration for longitudinal gestational monitoring.

References

[1] M. Feng, L. Wen, Z. Li, L. Qin, and X. Qi, "Fetal Weight Estimation Via Ultrasound Using Machine Learning," *IEEE Access*, DOI: 10.1109/ACCESS.2019.2933801, 2019.

[2] M. S. Alam, A. S. M. S. Ahamed, and O. Bappe, "A Comprehensive Approach to Fetal Health Prediction Using Machine Learning and Ensemble Models," in *Proc. 27th Int. Conf. Computer and Information Technology (ICCIT)*, Cox's Bazar, Bangladesh, Dec. 2024, pp. 2345–2350, doi: 10.1109/ICCIT64611.2024.11022031.

[3] M. F. Ukey, L. K. Alam, and V. Chauhan, "Fetal Health Classification and Visceral Fat Level Prediction using Gradient Boosting and Deep Learning Techniques," in *Proc. 2nd Int. Conf. Networking and Communications (ICNWC)*, 2024, doi: 10.1109/ICNWC60771.2024.10537331.

[4] S. A. M., S. H., C. Ramakrishnan, and T. V. Narmadha, "Effective Feature Technique for Heart Disease [FETUS] Prediction in Machine Learning," in *Proc. Int. Conf. Frontier Technologies and Solutions (ICFTS)*, 2015, doi: 10.1109/ICFTS62006.2023.10015655.

[5] N. Rahmayanti, H. Pradani, M. Pahlawan, and K. Vinarti, "Comparison of machine learning algorithms to classify fetal health using cardiotocogram data," *Procedia Computer Science*, vol. 197, pp. 162–171, 2022.

[6] H. Yasaswy and A. Singh, "A Machine Learning Based Prediction Model for Fetal Health Assessment," in *Frontiers of ICT in Healthcare: Proc. ICICT 2022*, Springer, 2023, pp. 219–230.

[7] A. Akbulut, E. Ertugrul, and V. Topcu, "Fetal health status prediction based on maternal clinical history using machine learning techniques," *Computer Methods and Programs in Biomedicine*, vol. 163, pp. 87–100, 2018.

[8] Y. Zhang and X. Zhao, "Fetal date assessment based on cardiotocography parameters using PCA and AdaBoost," in *Proc. 10th Int. Congress Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2017, pp. 1–6.

[9] Y. Gong *et al.*, "Fetal congenital heart disease echocardiogram screening based on DGACNN: Adversarial One-class classification combined with Video transfer learning," *IEEE Trans. Medical Imaging*, vol. 39, no. 4, pp. 1206–1222, 2020.

[10] F. A. Warrick, E. F. Hamilton, H. Penny, and R. E. Kearney, "Classification of Normal and Hypoxic Fetuses From Systems Modeling of Intrapartum Cardiotocography," *IEEE Trans. Biomedical Engineering*, vol. 57, no. 4, pp. 771–779, Apr. 2010.

[11] G. Georgoulas, D. Stylios, and P. Groumpos, "Predicting the risk of metabolic acidosis for newborns based on fetal heart rate signal classification using support vector machines," *IEEE Trans. Biomedical Engineering*, vol. 53, no. 5, pp. 875–884, May 2006.

[12] J. Spilka *et al.*, "Using nonlinear features for fetal heart rate classification," *Biomedical Signal Processing and Control*, vol. 7, no. 4, pp. 350–357, Jul. 2012.

[13] E. F. Hadlock, R. B. Harist, R. S. Sharman, R. L. Deter, and S. K. Park, "Estimation of fetal weight with the use of head, body, and femur measurements—a prospective study," *Am. J. Obstetrics and Gynecology*, vol. 151, no. 3, pp. 333–337, 1985.

[14] M. J. Shepard, V. A. Richards, R. J. Berkowitz, S. L. Warsof, and J. C. Hobbins, "An evaluation of two equations for predicting fetal weight by ultrasound," *Am. J. Obstetrics and Gynecology*, vol. 142, no. 1, pp. 47–54, 1982.

[15] M. J. Dudley, "A systematic review of the ultrasound estimation of fetal weight," *Ultrasound in Obstetrics and Gynecology*, vol. 25, no. 1, pp. 80–89, 2005.

[16] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[17] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[19] A. Hinchman Miell and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. 22nd Int. Conf. Machine Learning (ICML)*, pp. 625–632, 2005.

[20] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine Learning*, vol. 51, no. 2, pp. 181–207, 2003.

[21] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[22] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[23] H. Sahin and A. Subasi, "Classification of the cardiotocogram data for anticipation of fetal risks using machine learning techniques," *Applied Soft Computing*, vol. 33, pp. 231–238, Aug. 2015.

[24] L. Liu, I. A. Dekkers, Q. Tao, and H. J. Lamb, "Novel artificial neural network and linear regression based equation for estimating visceral adipose tissue volume," *Clinical Nutrition*, vol. 39, no. 10, pp. 3182–3188, 2020.

[25] S. Badshah *et al.*, "Risk factors for low birth weight in the public hospitals of Peshawar, NWFP-Pakistan," *BMC Public Health*, vol. 8, p. 197, 2008.

[26] H. G. Mengesha, A. D. Wunh, B. Woldemariyam, and S. D. Mengesha, "Low birth weight and macronutrients in Tigray, Northern Ethiopia: who are the mothers at risk?," *BMC Pediatrics*, vol. 17, no. 1, p. 148, 2017.

[27] T. Hoshiai, "The macrosomic fetus: a challenge in current obstetrics," *Acta Obstetricia et Gynecologica Scandinavica*, vol. 87, no. 2, pp. 134–145, 2008.

[28] L. M. Bernstein and P. M. Caldeyro, "Influence of fetal fat on the ultrasonic estimation of fetal weight in diabetic mothers," *Obstetrics and Gynecology*, vol. 79, no. 4, pp. 561–563, 1992.