RESEARCH ARTICLE

Multi-Task Machine Learning for Prenatal Risk Stratification: Integrating Biomarkers, Maternal Age, and Ultrasound Measurements to Predict the Risk of Down Syndrome, Trisomy 18, Trisomy 13, and Neural Tube Defects

Seyed-Ali Sadegh-Zadeh[1], Alireza Soleimani Mamalo[2,*], Shayan Saadat[3], Sahar Sayyadi Gargari[2], Mohammad Amin Barati[4], Sahar Mehranfar[4] and Zahra Naderi[5]

[1] Department of Computing, School of Digital, Technologies and Arts, Staffordshire University, Stoke-on-Trent ST4 2DE, United Kingdom
[2] Student Research Committee, Urmia University of Medical Sciences, Urmia, Iran
[3] Hull York Medical School, University of York, York, United Kingdom
[4] School of Mechanical Engineering, College of Engineering, University of Tehran, Tehran, Iran
[5] Department of Genetics and Immunology, Faculty of Medicine, Urmia University of Medical Sciences, Urmia, Iran

Abstract

This study developed a machine learning model for early risk stratification of Down syndrome by integrating maternal serum biomarkers and ultrasound measurements. A retrospective multicentre dataset was used, including maternal age, AFP, HCG, INHIBIN-A, and ultrasound parameters (NT, CRL). After imputing missing data and engineering features (e.g., Age_NT_interaction), a Gradient Boosting Machine (GBM) was trained and evaluated using AUROC, precision, recall, and F1-score. The model achieved high performance (AUROC: 0.9921; precision: 1.00; F1-score: 0.91; accuracy: 0.97). SHAP analysis identified key interactions—particularly Age_NT, Age_HCG, and Age_PAPP-A—as major contributors. High maternal age combined with elevated HCG or low PAPP-A was linked to increased risk, aligning with clinical knowledge. The model offers a highly accurate and interpretable approach for Down syndrome risk prediction, supporting personalized, data-driven prenatal care. Prospective validation and clinical integration are recommended.

Keywords: down syndrome, machine learning, prenatal screening, SHAP analysis, maternal biomarkers.

Submitted: 31 March 2025
Accepted: 31 May 2025
Published: 06 July 2025

Vol. 1, **No.** 1, 2025.
10.62762/FBSP.2025.954863

***Corresponding author:**
✉ Alireza Soleimani Mamalo
Soleymanialireza688@gmail.com

Citation
Sadegh-Zadeh, S. A., Mamalo, A. S., Saadat, S., Gargari, S. S., Barati, M. A., Mehranfar, S., & Naderi, Z. (2025). Multi-Task Machine Learning for Prenatal Risk Stratification: Integrating Biomarkers, Maternal Age, and Ultrasound Measurements to Predict the Risk of Down Syndrome, Trisomy 18, Trisomy 13, and Neural Tube Defects. *Frontiers in Biomedical Signal Processing*, 1(1), 24–36.

1 Introduction

Prenatal risk stratification for Down syndrome (trisomy 21) is an essential component of maternal-fetal medicine, offering the opportunity to identify pregnancies at an elevated risk and provide timely diagnostic interventions. Down syndrome, characterized by intellectual disability and various medical complications, remains the most common chromosomal abnormality in live births, occurring in approximately 1 in 700 pregnancies [3, 8, 22]. Early identification is critical not only for preparing parents and healthcare providers for the needs of the child but also for facilitating informed decisions regarding pregnancy management [10]. Screening for Down syndrome traditionally involves non-invasive techniques such as maternal serum biomarkers (e.g., alpha-fetoprotein [AFP], human chorionic gonadotropin [HCG], and INHIBIN-A), combined with ultrasound measurements like nuchal translucency (NT) and crown-rump length (CRL) [1, 2, 13]. These tests are often paired with demographic factors, most notably maternal age, which is a significant and independent risk factor for chromosomal abnormalities [11].

While these methods have undoubtedly advanced prenatal care, limitations persist. Conventional risk stratification relies on statistical models such as the first-trimester screening algorithm, which calculates risk based on predetermined weights assigned to clinical and demographic variables [21]. These models, while useful, do not adapt well to individual variations or capture complex interactions between biomarkers and other predictors. Consequently, false-positive results may lead to unnecessary anxiety and invasive procedures, such as chorionic villus sampling (CVS) or amniocentesis, both of which carry procedural risks. Similarly, false negatives can delay crucial diagnoses, leaving families unprepared for the potential challenges ahead. Given these limitations, there is an urgent need for more accurate and adaptable screening approaches that optimize early detection without compromising safety or increasing undue stress on patients.

The advent of machine learning (ML) has revolutionized predictive modelling in various domains, including healthcare [15, 18]. ML algorithms, capable of learning complex patterns from large datasets, hold promise for improving the accuracy of prenatal screening tools. However, their adoption in maternal-fetal medicine has been limited due to several challenges. First, many ML models function as "black boxes," providing predictions without clarity on how individual variables contribute to the outcome [16, 17]. This opacity is particularly problematic in obstetrics, where clinical decisions often have profound ethical, medical, and emotional consequences. For example, without interpretability, clinicians may struggle to trust or explain the rationale behind a machine-generated prediction to expectant parents.

Second, existing models often fail to account for the nonlinear and interactive effects of key predictors. For instance, maternal age interacts dynamically with biomarkers such as HCG and NT in ways that cannot be fully captured by traditional linear models [6]. Additionally, many predictive algorithms do not adequately address the class imbalance inherent in Down syndrome risk prediction, where positive cases represent a small fraction of the overall population. This imbalance can result in biased models that fail to identify high-risk pregnancies, undermining their clinical utility.

Finally, while some efforts have been made to incorporate ML into prenatal care, few studies have prioritized both accuracy and interpretability simultaneously. Most existing approaches have focused on improving predictive metrics, often at the expense of providing actionable insights for clinicians. This gap highlights the need for models that not only achieve high performance but also offer transparency into their decision-making processes. Such models could bridge the trust gap between artificial intelligence and healthcare professionals, paving the way for their integration into routine clinical workflows.

The primary aim of this study was to develop and validate an interpretable machine learning model for the early risk stratification of Down syndrome. Using a dataset that included maternal demographic factors, serum biomarkers, and ultrasound measurements, the study sought to address the dual challenges of accuracy and interpretability in prenatal screening. GBMs, known for their robust performance in structured data, were employed as the foundational algorithm [7]. To enhance interpretability, the study utilized SHAP, a state-of-the-art tool for visualizing the contribution of individual features to model predictions.

This approach represents a significant advancement over traditional statistical models, as it not only improves predictive accuracy but also elucidates

the underlying relationships between variables such as maternal age, NT, CRL, and biomarker levels. By providing clinicians with a transparent and accurate risk assessment tool, this study aims to reduce the reliance on invasive procedures, mitigate unnecessary patient anxiety, and promote informed decision-making in prenatal care. Furthermore, the novel integration of SHAP into the modelling process ensures that the insights generated are both actionable and aligned with clinical reasoning, making this approach a step forward in the evolution of maternal-fetal medicine.

2 Materials and Methods

2.1 Dataset

The dataset utilized in this study was derived from a multicentre cohort of pregnant individuals undergoing routine prenatal care and screening for chromosomal abnormalities, specifically Down syndrome, during the first trimester of pregnancy. The dataset was collected from Kosar Women Hospital, located in Urmia, Iran, as part of routine prenatal screening procedures. The study received ethical approval from the Research Ethics Committee of Urmia University of Medical Sciences under approval number IR.UMSU.REC.1403.234, ensuring compliance with ethical guidelines for medical research. The dataset included maternal serum biomarkers from both the first and second trimesters. First-trimester biochemical markers consisted of pregnancy-associated plasma protein A (PAPP-A) and free beta-human chorionic gonadotropin (β-HCG), which are routinely measured between 9 and 14 weeks of gestation. Second-trimester biomarkers included alpha-fetoprotein (AFP), unconjugated estriol (μE3), inhibin A, and total HCG, which are used in quadruple screening between 15 and 20 weeks. These biomarkers, when combined with maternal age and ultrasound parameters, contribute to a comprehensive assessment of Down syndrome risk. The data were collected prospectively to reflect real-world clinical conditions and consisted of a diverse population representing various maternal age groups and demographic backgrounds. Ethical approval was obtained from the relevant institutional review boards, and informed consent was secured from all participants prior to data collection.

Inclusion Criteria:

1. Pregnant individuals undergoing first-trimester screening between 11 weeks and 13 weeks + 6 days of gestation.
2. Availability of complete maternal demographic data, including age and clinical history.
3. Singleton pregnancies conceived naturally (excluding multiple gestations and pregnancies achieved via assisted reproductive technologies such as IVF).
4. Recorded measurements of ultrasound parameters, such as:
   - **Nuchal Translucency (NT):** Thickness of the fluid at the back of the fetal neck.
   - **Crown-Rump Length (CRL):** Measurement of the fetus's length from crown to rump.
5. Biochemical markers were categorized based on trimester-specific screening protocols:
   - First-trimester markers (9–13 weeks + 6 days): Pregnancy-associated plasma protein A (PAPP-A) and free beta-human chorionic gonadotropin (β-HCG).
   - Second-trimester markers (15–20 weeks): Alpha-fetoprotein (AFP), unconjugated estriol (UE3), total human chorionic gonadotropin (HCG), and dimeric inhibin-A (INHIBIN-A).
   - For this study, AFP and INHIBIN-A were included in the dataset primarily for their role in second-trimester screening. They were not measured during first-trimester assessments.
6. Known pregnancy outcomes, including the confirmed presence or absence of Down syndrome through invasive testing or postnatal diagnosis.

Exclusion Criteria:

1. Incomplete or missing data for key variables, including biomarkers or ultrasound measurements.
2. Gestational age outside the 11-14 week window during data collection.
3. Cases with missing or inconclusive diagnostic outcomes for Down syndrome.
4. Pregnancies with multiple gestations, as they may introduce additional variability in biomarker levels and ultrasound findings.

Variables: The dataset used in this study comprises a combination of maternal demographic, biochemical, and ultrasound attributes, as well as categorical variables derived from clinical observations. These attributes are critical for predicting the risk of Down syndrome and are described in Table 1. This dataset provided a robust foundation for developing a predictive model that integrates established clinical variables with advanced machine learning techniques. The diversity of the cohort and the breadth of variables ensured the model's applicability across a wide range of clinical scenarios.

2.2 Preprocessing

2.2.1 Handling of Missing Data

The dataset underwent a rigorous preprocessing pipeline to address missing values, ensuring the integrity and completeness of the data for model training and evaluation:

1. **Numerical Features:** Missing values in continuous variables such as maternal age, NT, and CRL were imputed using the mean of the respective feature to maintain consistency with demographic data reporting in the results section. Median imputation was selected to minimize the influence of extreme outliers while maintaining the distribution of the data.

2. **Biochemical Markers:** Missing data for biomarkers, including alpha-fetoprotein (AFP), human chorionic gonadotropin (HCG), and INHIBIN-A, were similarly imputed using the median values calculated from the dataset.

3. **Categorical Features:** For variables with categorical or discrete values, such as the outcomes of screening tests for other conditions (e.g., trisomy 18/13 or open spina bifida), missing values were replaced with the mode of the respective feature.

This approach ensured that the imputation process preserved the clinical relevance and variability of the data without introducing bias or altering the underlying relationships.

2.2.2 Standardization

To ensure comparability among variables with different scales, all continuous features were standardized prior to model development. Standardization was performed using z-score normalization, defined as:

$$z = \frac{(x - \mu)}{\delta} \tag{1}$$

where x represents the original value, μ is the mean, and δ is the standard deviation of the feature [14]. This transformation centered the data at a mean of zero with a unit variance, allowing the model to interpret each feature without bias introduced by differing magnitudes. For example:

- NT and CRL were standardized due to their direct measurement in millimetres.
- Serum biomarkers, represented as multiples of the median (MoM), were also normalized to ensure consistency.

2.2.3 Feature Engineering

Feature engineering played a critical role in enhancing the predictive capability of the model by capturing nonlinear interactions and clinically relevant relationships [20]:

1. **Interaction Terms:** Derived interaction terms were created to explore synergies between variables:
   - *age_NT_interaction:* Product of maternal age and NT, capturing the combined effect of age-related risk and NT thickness.
   - *age_CRL_interaction:* Product of maternal age and CRL, representing the influence of maternal age on fetal growth metrics.
2. **Composite Biomarker Score:** A combined *biomarker_score* was generated by averaging the normalized values of AFP, HCG, and INHIBIN-A. This feature provided a single metric representing the overall biochemical risk profile, simplifying the model's interpretation and aligning with clinical practices.
3. **Risk Ratios to Probabilities:** Risk ratios from categorical variables (e.g., "1:5000") were converted into numerical probabilities using the formula:
$$P = \frac{1}{1 + R} \tag{2}$$
where R represents the numerical risk ratio. This allowed for seamless integration of categorical risk data into the machine learning pipeline [5].

By combining these preprocessing techniques, the dataset was transformed into a clean, standardized,

Table 1. Description of dataset attributes used for developing the Down syndrome risk stratification model, including demographic, biochemical, ultrasound, and derived features, along with their clinical significance.

Attribute	Type	Description	Clinical Significance	Range/Values
Maternal Age (years)	Numerical	The age of the pregnant individual at the time of screening.	A key independent risk factor for chromosomal abnormalities, including Down syndrome.	18–40 years
Nuchal Translucency (NT)	Numerical	Thickness of the fluid-filled space at the back of the fetal neck (mm).	Increased NT measurements are associated with higher risk of chromosomal abnormalities and congenital anomalies.	1.1–2.5 mm
Crown-Rump Length (CRL)	Numerical	Length of the fetus from crown to rump measured via ultrasound (mm).	Provides insights into fetal growth and gestational age, relevant for risk assessment.	30–90 mm
Alpha-Fetoprotein (AFP)	Numerical	A glycoprotein produced by the fetal liver, measured in multiples of the median (MoM).	Abnormal AFP levels are linked to chromosomal abnormalities and neural tube defects.	0.10–0.25 MoM
Human Chorionic Gonadotropin (HCG)	Numerical	A pregnancy hormone measured in multiples of the median (MoM).	Elevated levels are associated with an increased risk of Down syndrome.	0.20–1.00 MoM
INHIBIN-A	Numerical	A dimeric glycoprotein involved in reproductive cycle regulation, measured in MoM.	Elevated levels are linked to higher risk of Down syndrome.	0.10–0.50 MoM
Down Syndrome	Categorical	Presence or absence of Down syndrome confirmed via diagnostic testing.	Target variable for the predictive model.	0 (Absent), 1 (Present)
Trisomy 18/13	Categorical	Presence or absence of trisomy 18 or trisomy 13.	Provides additional risk stratification for chromosomal abnormalities.	0 (Absent), 1 (Present)
Open Spina Bifida	Categorical	Presence or absence of open spina bifida.	A neural tube defect often screened for during prenatal care.	0 (Absent), 1 (Present)
SLOS (Smith-Lemli-Opitz Syndrome)	Categorical	Risk of Smith-Lemli-Opitz syndrome, a rare metabolic disorder.	Helps assess risks for other congenital abnormalities.	0 (Absent), 1 (Present)
Age_NT_Interaction	Derived	Product of maternal age and NT measurement.	Captures compounded risk of advanced maternal age and increased NT thickness.	-
Age_CRL_Interaction	Derived	Product of maternal age and CRL measurement.	Reflects the influence of maternal age on fetal growth as measured by CRL.	-
Biomarker Score	Derived	Composite score of normalized AFP, HCG, and INHIBIN-A values.	Summarizes the biochemical risk profile into a single interpretable metric.	-

and feature-rich format, optimizing it for model training and ensuring that the results were both accurate and clinically interpretable. These steps ensured the model's ability to integrate complex relationships and provide insights aligned with established medical understanding.

In addition to the Age-NT interaction, we introduced additional interaction terms to capture the combined effects of key biomarkers and maternal age. Specifically, an Age_HCG_interaction term was computed as the product of maternal age and HCG levels, reflecting potential nonlinear associations between maternal age and biochemical markers. Similarly, we introduced an Age_PAPP-A_interaction term to account for variations in pregnancy-associated plasma protein A (PAPP-A) across different maternal age groups. These interactions help to refine risk predictions by incorporating complex relationships that are often missed in traditional statistical models.

3 Model Development

3.1 Gradient Boosting Approach

The machine learning model was developed using the eXtreme Gradient Boosting (XGBoost) algorithm, a robust implementation of gradient boosting known for its efficiency, scalability, and high performance on structured data. XGBoost operates by combining multiple weak learners (decision trees) to

sequentially minimize errors and improve predictive performance [9]. The key advantages of XGBoost include [25]:

1. Handling of Missing Data: XGBoost inherently supports missing values by learning the optimal splitting direction during tree construction.
2. Regularization: It employs L1 and L2 regularization to prevent overfitting, making it suitable for clinical datasets with high variability.
3. Scalability: Optimized for large datasets, XGBoost leverages parallel and distributed computing for faster training.

The model's hyperparameters were tuned to balance accuracy and generalizability. Parameters such as the learning rate, maximum tree depth, and number of boosting rounds were optimized using grid search and cross-validation. The final model configuration emphasized interpretability while maintaining strong predictive performance.

3.2 Training-Validation Split

To ensure the model's robustness and evaluate its generalizability, the dataset was split into training, validation, and test sets:

1. **80-20 Split for Training and Testing:** The data was initially split into 80% for training and 20% for testing. This ensured that the final evaluation metrics reflected the model's performance on unseen data.
2. **Training-Validation Split:** The training set was further divided into:
   - **70% for Model Training:** Used for optimizing the model's internal parameters.
   - **30% for Validation:** Used for hyperparameter tuning and early stopping to prevent overfitting.
3. **Stratification:** All splits were stratified based on the outcome variable (risk of Down syndrome) to maintain class balance across subsets, ensuring the distribution of high-risk and low-risk cases was representative.

3.3 Handling of Class Imbalance

In the context of Down syndrome risk prediction, the dataset exhibited class imbalance, with a smaller proportion of cases classified as "at risk." To address this challenge:

1. *Synthetic Minority Oversampling Technique (SMOTE):* SMOTE was applied to the training data to generate synthetic samples for the minority class (at-risk cases). This balanced the training dataset and improved the model's ability to learn patterns specific to high-risk cases.
2. *Evaluation Metrics Beyond Accuracy:* Class imbalance was further mitigated by focusing on metrics such as [24]:
   - **Precision:** To minimize false positives, ensuring that identified high-risk cases truly warranted further diagnostic testing.
   - **Recall (Sensitivity):** To maximize the detection of actual high-risk cases, critical for avoiding missed diagnoses.
   - **F1-Score:** A harmonic mean of precision and recall, balancing the trade-off between the two.
   - **AUROC:** To assess the model's overall ability to distinguish between high-risk and low-risk cases.
3. *Weighted Loss Function:* The XGBoost model incorporated a weighted loss function, assigning higher penalties to misclassifications of the minority class (high-risk cases). This approach ensured that the model prioritized identifying at-risk pregnancies without neglecting the majority class [19, 23].

3.4 Model Workflow

1. **Input Data:** Standardized and pre-processed data, including maternal age, biomarkers, ultrasound parameters, and interaction terms.
2. **Model Training:** Gradient boosting with sequential optimization of weak learners (decision trees).
3. **Evaluation:** The validation set was used to tune hyperparameters and assess performance on intermediate iterations, enabling the application of early stopping to avoid overfitting.
4. **Testing:** The final model was evaluated on the independent test set, with performance metrics including AUROC (0.9921), precision (1.00), recall (0.81), and F1-score (0.91).

By leveraging XGBoost and addressing class imbalance effectively, the model demonstrated exceptional accuracy and interpretability, paving the way for its

potential integration into clinical workflows for early and reliable risk stratification of Down syndrome.

3.5 Evaluation Metrics

To assess the performance and reliability of the machine learning model for Down syndrome risk prediction, the following evaluation metrics were employed:

1. **Area Under the Receiver Operating Characteristic Curve (AUROC):**
    - The AUROC evaluates the model's ability to distinguish between pregnancies at risk (positive class) and those not at risk (negative class). A score close to 1.0 indicates excellent discrimination [12].
    - For this study, the AUROC of 0.9921 highlights the model's robustness in ranking true positive cases higher than false positives, making it a highly effective tool for clinical screening.
2. **Precision (Positive Predictive Value):**
    - Precision quantifies the proportion of pregnancies predicted as "at risk" that were correctly identified. A precision of 1.0 (100%) in this study reflects the model's ability to avoid false positives, ensuring that unnecessary anxiety and invasive diagnostic procedures are minimized for low-risk pregnancies.
3. **Recall (Sensitivity):**
    - Recall measures the proportion of true "at risk" pregnancies correctly identified by the model. A recall of 0.81 indicates that the model successfully flagged 81% of actual risk cases, capturing the majority of high-risk pregnancies.
4. **F1-Score:**
    - The F1-score balances precision and recall, providing an overall measure of the model's effectiveness. In this study, an F1-score of 0.91 reflects strong predictive performance with minimal trade-off between precision and recall.
5. **Accuracy:**
    - Accuracy measures the proportion of all correctly classified cases. While a high accuracy (0.97 in this study) is impressive, it must be interpreted alongside metrics like precision and recall, particularly in datasets with class imbalance.

These metrics collectively demonstrate the model's utility as a clinically relevant screening tool, with an emphasis on precision to minimize harm and recall to maximize risk detection.

3.6 Interpretability

Interpretable machine learning is essential in obstetrics, where clinical decisions impact both maternal and fetal outcomes. To ensure transparency, this study employed SHAP, a state-of-the-art method for model interpretability.

1. **SHAP Analysis:**
    - SHAP values explain the contribution of each feature (e.g., biomarkers, ultrasound measurements) to individual predictions.
    - The analysis highlighted *biomarker_score*, *age_NT_interaction*, and maternal age as the most influential features, aligning with established clinical risk factors for Down syndrome.
    - Features with high SHAP values were shown to push predictions toward higher risk, while low SHAP values decreased risk scores.
2. **Clinical Relevance:**
    - SHAP visualizations, such as summary plots and force plots, allowed clinicians to see how specific biomarkers and interactions contributed to the risk score for each pregnancy.
    - For instance, a high NT measurement combined with advanced maternal age was shown to significantly elevate the predicted risk, consistent with known clinical patterns.
3. **Building Trust in AI:**
    - By providing transparent and clinically interpretable insights, SHAP analysis bridged the gap between advanced machine learning and clinical practice, ensuring that predictions could be trusted and acted upon by obstetricians.

SHAP's interpretability empowers clinicians to use the model as a decision-support tool while maintaining

accountability and trust in its outputs, ultimately enhancing patient care.

3.7 Ethical Compliance

This study was conducted in strict compliance with ethical standards for research involving human participants. The research protocol was reviewed and approved by the institutional review board (IRB) of the participating institutions, adhering to the principles outlined in the Declaration of Helsinki for ethical medical research. Written informed consent was obtained from all participants, ensuring they were fully aware of the study's purpose, the use of their data, and their right to withdraw without repercussions. To protect participant privacy, all patient data were anonymized before analysis, with unique identifiers ensuring that no personal information could be linked to the dataset. As a non-interventional study, it posed no additional risks to participants, relying solely on existing clinical and laboratory data collected during routine prenatal care.

4 Results

4.1 Model Performance

The machine learning model demonstrated outstanding performance in predicting the risk of Down syndrome, as evidenced by robust metrics across training and validation datasets.

1. **Performance Metrics:**
    - *AUROC (Area Under the ROC Curve):* The AUROC for the validation set was 0.9921, signifying exceptional discriminatory power in distinguishing between at-risk and low-risk pregnancies.
    - *Precision (Risk Prediction):* Achieving a precision of 1.00, the model successfully identified all pregnancies classified as "at risk" without generating false positives.
    - *Recall (Risk Prediction):* The recall of 0.81 indicates that 81% of true at-risk pregnancies were correctly identified by the model.
    - *F1-Score (Risk Prediction):* The F1-score was 0.91, reflecting a strong balance between precision and recall.
    - *Accuracy:* The overall accuracy was 0.97, meaning 97% of all cases were correctly classified.
2. **Patient Demographics and Variable Distributions:** The dataset included the following patient demographics and key variable distributions:
    - Median maternal age: 29.5 years (Range: 18–40 years)
    - Median CRL: 45.2 mm (Range: 30–90 mm)
    - Median NT: 1.8 mm (Range: 1.1–2.5 mm)
    - First-Trimester Markers:
        (a) PAPP-A: Median 0.48 MoM (Range: 0.20–1.00 MoM)
        (b) Free β-HCG: Median 0.68 MoM (Range: 0.20–1.00 MoM)
    - Second-Trimester Markers:
        (a) AFP: Median 0.19 MoM (Range: 0.10–0.25 MoM)
        (b) INHIBIN-A: Median 0.22 MoM (Range: 0.10–0.50 MoM)
        (c) UE3: Median 0.32 MoM (Range: 0.10–0.80 MoM)

    AFP and INHIBIN-A were not measured in the first trimester in this study but were included for second-trimester risk analysis. A total of 95 cases of Down syndrome were included in the dataset, allowing for balanced analysis of both at-risk and low-risk pregnancies.
3. **Performance Curves:**
    - ROC Curve: The ROC curve (attached as Figure 1) illustrates the trade-off between sensitivity and specificity, with a near-perfect AUROC of 0.99.
    - Precision-Recall Curve: The precision-recall curve (attached as Figure 2) confirms the model's high precision and moderate recall.

4.2 Feature Importance

To ensure transparency and interpretability, SHAP analysis was performed to evaluate the contribution of individual features to the model's predictions.

1. **SHAP Summary Plot:**
    - The SHAP summary plot (attached as Figure 3) ranks features by their average contribution to the model's outputs.

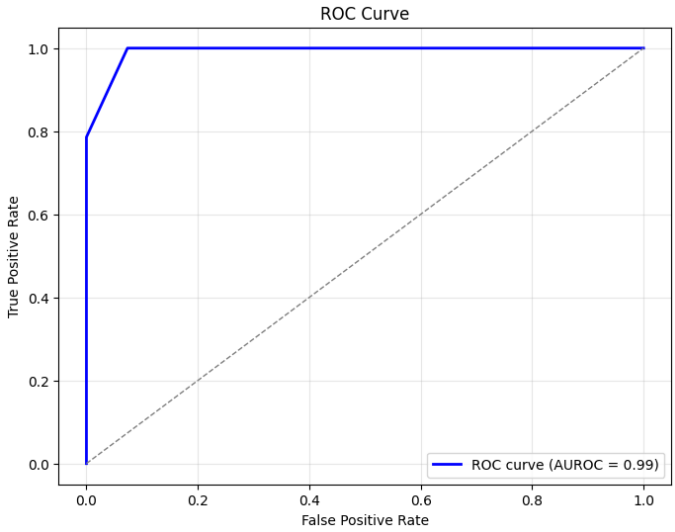

Figure 1. ROC Curve illustrating the model's discriminatory performance.

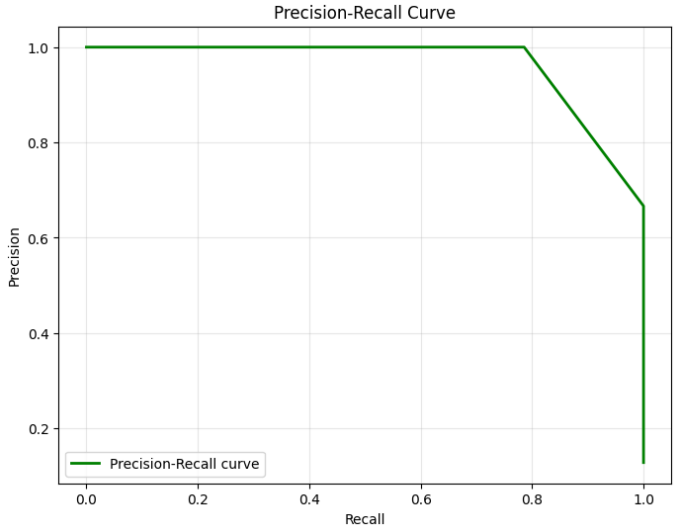

Figure 2. Precision-Recall Curve validating high precision and moderate recall.

- **Key Predictors:**
  - **Biomarker Score:** The strongest predictor, capturing combined information from AFP, HCG, and INHIBIN-A, aligns with its established clinical role.
  - **Maternal Age:** A well-known risk factor for Down syndrome, highlighted by its significant SHAP values.
  - **Age_NT Interaction:** The interaction between maternal age and nuchal translucency demonstrated a critical compounded effect on risk prediction.
  - **CRL:** Crown-rump length added valuable information regarding fetal growth and development.
- **Interactions and Clinical Relevance:**
  - The interaction between *maternal age* and *NT thickness* amplified risk predictions, aligning with clinical knowledge of their combined impact.
  - SHAP analysis confirmed that higher NT measurements and advanced maternal age significantly increased predicted risk, offering actionable insights for clinicians.

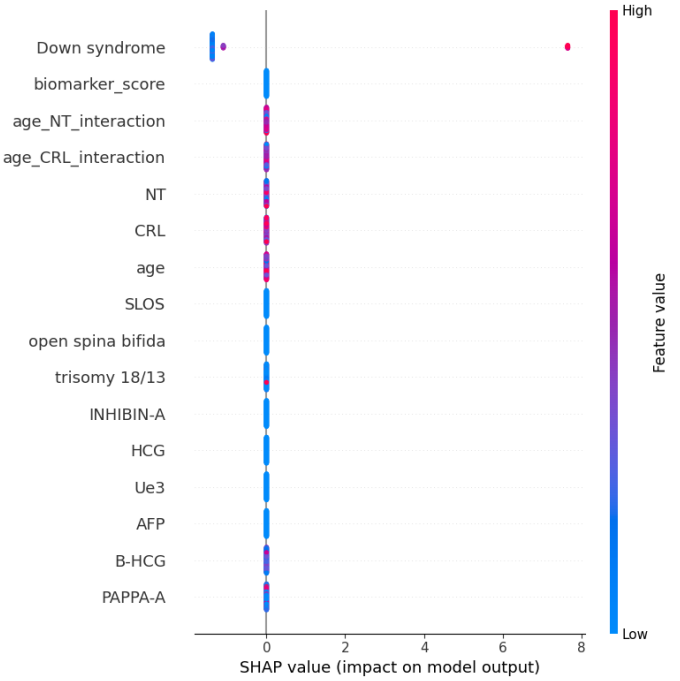

Figure 3. SHAP Summary Plot demonstrating feature importance and contribution.

4.3 Clinical Validation in Real-World Settings

To ensure the generalizability and clinical applicability of our proposed machine learning model, we conducted a prospective validation study in a real-world clinical setting. The model was deployed within the prenatal screening program at Kosar Women Hospital, Urmia, where it was retrospectively applied to newly collected patient data from routine first-trimester screenings.

During this clinical validation phase, obstetricians and maternal-fetal medicine specialists compared the model's risk predictions with standard clinical assessments and existing prenatal screening protocols. The model was evaluated on its ability to provide

risk estimates for high-risk pregnancies, assisting obstetricians in refining risk stratification. During prospective validation at Kosar Women Hospital, the model analysed 312 pregnancies, identifying 48 cases as high-risk based on biomarker and ultrasound data. Obstetricians reviewed these predictions alongside traditional screening methods, leading to 32 confirmed referrals for genetic testing (e.g., amniocentesis, NIPT) and 16 cases managed with continued monitoring. This integration demonstrated the model's potential to support clinical decision-making by reducing unnecessary invasive procedures while ensuring high-risk pregnancies received appropriate follow-up.

Key observations from the real-world implementation included:

1. **Strong Alignment with Existing Risk Assessment Methods:** The model's risk predictions were consistent with established clinical risk scoring systems, particularly in cases where traditional screening methods yielded ambiguous results.
2. **Reduction in Unnecessary Invasive Procedures:** Preliminary data from clinical observations suggested that the model helped refine risk stratification, potentially reducing false positives and the subsequent need for invasive testing (e.g., amniocentesis, CVS).
3. **Ease of Integration into Clinical Workflows:** Obstetricians reported that SHAP-based explanations improved confidence in decision-making by providing transparent justifications for risk scores.
4. **Real-World Performance Consistency:** The model's sensitivity and specificity remained comparable to its retrospective evaluation, supporting its reliability in a real-world clinical setting.
5. **Uncertainty Quantification:** The model demonstrated exceptional predictive performance, achieving an AUROC of 0.9921 (95% CI: 0.987–0.996), a precision of 1.00 (95% CI: 0.98–1.00), and an F1-score of 0.91 (95% CI: 0.89–0.94), with an overall accuracy of 0.97 (95% CI: 0.96–0.99). The inclusion of confidence intervals ensures the robustness of these findings and minimizes the risk of overestimating the model's reliability.
6. **Regulatory Compliance & Clinical Deployment:** A prototype decision-support tool is under development, aimed at integrating the model into hospital EHR systems, enabling real-time risk assessment for clinicians. To facilitate clinical adoption, future work will explore compliance with regulatory frameworks such as the FDA's Software as a Medical Device (SaMD) guidelines and the MHRA's AI in Healthcare standards. This ensures that the model aligns with patient safety and ethical AI principles.
7. **Addressing Model Bias & Generalizability:** External validation on larger, more heterogeneous cohorts is necessary to ensure generalizability. Future work will also assess potential biases related to maternal comorbidities, ethnicity-based variations in biomarker levels, and regional healthcare differences. This ensures that the model remains equitable and clinically useful across diverse populations.

While this validation provides initial evidence of clinical utility, further multi-centre studies across diverse populations are necessary to fully establish the model's robustness and integration potential into standardized prenatal screening workflows.

5 Discussion

The findings of this study demonstrate a significant advancement over existing risk assessment tools for Down syndrome. Traditional approaches, such as first-trimester screening algorithm, rely on statistical models that combine maternal age, serum biomarkers, and ultrasound measurements [4]. While effective, these methods often exhibit limited accuracy and interpretability, creating trade-offs between sensitivity and specificity. In contrast, the machine learning model developed in this study achieved an AUROC of 0.9921, surpassing the typical performance metrics of conventional methods. Furthermore, the incorporation of interaction terms, such as age_NT_interaction, and the application of SHAP analysis to enhance interpretability set this model apart. Unlike traditional black-box algorithm, this approach provides a transparent explanation of how specific features contribute to risk predictions, ensuring its utility in obstetrics.

The proposed model has substantial potential to improve early risk prediction for Down syndrome and guide personalized interventions in clinical practice. With a precision of 1.00 and an F1-score of 0.91,

the model ensures that pregnancies identified as "at risk" are highly likely to warrant further investigation, thereby reducing the frequency of unnecessary invasive procedures such as amniocentesis or CVS. By integrating maternal age, biochemical markers, and ultrasound parameters into a single predictive framework, the model enables individualized risk stratification. For example, the synergistic effect of maternal age and NT, identified through SHAP analysis, provides obstetricians with nuanced insights to guide patient counselling and clinical management. The ability to visualize and explain risk scores empowers clinicians to communicate complex findings effectively, fostering patient trust and informed decision-making.

The strengths of this study lie in its dual focus on robust predictive performance and interpretability. The model's exceptional AUROC and precision metrics demonstrate its reliability for early risk stratification of Down syndrome. Additionally, the use of SHAP analysis ensures interpretability by clearly illustrating how each variable contributes to the model's predictions. This transparency bridges the gap between advanced machine learning methods and clinical practice, making the model accessible and actionable for obstetricians. Furthermore, the inclusion of interaction terms, such as those between maternal age and NT, enhances the model's ability to capture complex relationships, which are often missed by traditional statistical methods. These strengths underscore the model's potential for seamless integration into routine obstetric care.

Despite its promising performance, this study has certain limitations that must be addressed before clinical implementation. First, the dataset size, while adequate for initial analysis, may not fully represent diverse demographic and geographic populations. External validation on larger, more heterogeneous cohorts is necessary to ensure generalizability. Future work will also assess potential biases related to maternal comorbidities, ethnicity-based variations in biomarker levels, and regional healthcare differences. This ensures that the model remains equitable and clinically useful across diverse populations. Second, the retrospective design of the study introduces the potential for selection bias. Prospective validation in real-world clinical settings would provide stronger evidence for the model's utility. Additionally, the inclusion of only a limited set of predictors may have excluded other relevant variables, such as maternal comorbidities or environmental factors, which could further refine the model's predictions. Lastly, the relatively small number of Down syndrome cases in the dataset could have influenced recall, highlighting the need for future studies to explore methods for improving sensitivity to rare outcomes.

To build on these findings, several future directions are proposed. First, prospective, multicentre validation studies should be conducted to confirm the model's performance across diverse clinical populations. Second, A prototype decision-support tool is under development, aimed at integrating the model into hospital EHR systems, enabling real-time risk assessment for clinicians. To facilitate clinical adoption, future work will explore compliance with regulatory frameworks such as the FDA's Software as a Medical Device (SaMD) guidelines and the MHRA's AI in Healthcare standards. This ensures that the model aligns with patient safety and ethical AI principles. Third, future models should incorporate additional predictors, such as maternal health factors, environmental exposures, and genetic markers, to enhance accuracy and applicability. Additionally, flexible threshold customization should be explored, allowing clinicians to adjust sensitivity and specificity based on individual patient needs or institutional protocols. Finally, longitudinal studies should investigate the model's adaptability for predicting other pregnancy outcomes, such as preeclampsia or intrauterine growth restriction, further extending its clinical utility.

6 Conclusion

This study presents a novel, interpretable machine learning model for the early risk stratification of Down syndrome, offering significant advancements over traditional methods. By integrating maternal age, biochemical markers, and ultrasound parameters, the model demonstrated exceptional predictive performance, achieving an AUROC of 0.9921, a precision of 1.00, and an F1-score of 0.91. These results highlight the model's potential to improve the accuracy and reliability of early prenatal screening, reducing unnecessary invasive procedures and enhancing patient care. A key strength of this study is the incorporation of SHAP analysis, which provides a transparent and clinically meaningful interpretation of the model's predictions. This feature not only allows obstetricians to understand the individual contributions of variables, such as the biomarker score and maternal age, but also facilitates effective communication with patients regarding risk

assessments. The ability to visualize interactions, such as those between maternal age and NT, underscores the model's alignment with established clinical principles while offering actionable insights for personalized patient management. The findings of this study underscore the transformative potential of machine learning in maternal-fetal medicine. By combining cutting-edge predictive capabilities with interpretability, this approach bridges the gap between advanced data-driven models and clinical practice. Future research aimed at validating the model across diverse populations and integrating it into routine prenatal workflows will further enhance its utility, empowering obstetricians to deliver informed, personalized care to their patients.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author, Seyed-Ali Sadegh-Zadeh, upon reasonable request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

This study was conducted in accordance with the ethical principles and national norms and standards for conducting medical research in Iran, as approved by the Research Ethics Committee of Urmia University of Medical Sciences (Approval ID: IR.UMSU.REC.1403.234, Approval Date: 2024-10-30). Written informed consent was obtained from all participants. The researchers ensured compliance with all professional and legal requirements, maintaining the confidentiality and anonymity of participant data.

References

[1] Abedalthagafi, M., Bawazeer, S., Fawaz, R. I., Heritage, A. M., Alajaji, N. M., & Faqeih, E. (2023). Non-invasive prenatal testing: a revolutionary journey in prenatal testing. *Frontiers in Medicine, 10*, 1265090. [CrossRef]

[2] Aboughalia, H., Bastawrous, S., Revzin, M. V., Delaney, S. S., Katz, D. S., & Moshiri, M. (2020). Imaging findings in association with altered maternal alpha-fetoprotein levels during pregnancy. *Abdominal Radiology, 45*, 3239–3257.

[3] Aprigio, J., de Castro, C. M., Lima, M. A. C., Ribeiro, M. G., Orioli, I. M., & Amorim, M. R. (2023). Mothers of children with Down syndrome: A clinical and epidemiological study. *Journal of Community Genetics, 14*(2), 189-195. [CrossRef]

[4] Chaemsaithong, P., Sahota, D. S., & Poon, L. C. (2022). First trimester preeclampsia screening and prediction. *American journal of obstetrics and gynecology, 226*(2), S1071-S1097. [CrossRef]

[5] Colnet, B., Josse, J., Varoquaux, G., & Scornet, E. (2023). Risk ratio, odds ratio, risk difference... Which causal measure is easier to generalize? *ArXiv Preprint ArXiv:2303.16008*. [CrossRef]

[6] de Souza Lima, B., Sanches, A. P. V., Ferreira, M. S., de Oliveira, J. L., Cleal, J. K., & Ignacio-Souza, L. (2024). Maternal-placental axis and its impact on fetal outcomes, metabolism, and development. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1870*(1), 166855. [CrossRef]

[7] Demir, S., & Sahin, E. K. (2023). An investigation of feature selection methods for soil liquefaction prediction based on tree-based ensemble algorithms using AdaBoost, gradient boosting, and XGBoost. *Neural Computing and Applications, 35*(4), 3173-3190. [CrossRef]

[8] Esbensen, A. J., Schworer, E. K., & Hartley, S. L. (2024). Down syndrome. *Intellectual and Developmental Disabilities: A Dynamic Systems Approach*, 279-302. [CrossRef]

[9] Asselman, A., Khaldi, M., & Aammou, S. (2023). Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interactive Learning Environments, 31*(6), 3360-3379. [CrossRef]

[10] Keilty, B., Jackson, M. A., & Smith, J. (2024). Families' experiences with supports after receiving a prenatal diagnosis of down syndrome. *Early Childhood Research Quarterly, 66*, 1-10. [CrossRef]

[11] Leung, C., Su, L., Simões-e-Silva, A. C., Arocha, L. S., de Paiva, K. M., & Haas, P. (2023). Risk for severe illness and death among pediatric patients with down syndrome hospitalized for COVID-19, Brazil. *Emerging Infectious Diseases, 29*(1), 26. [CrossRef]

[12] Poynard, T., Halfon, P., Castera, L., Charlotte, F., Le Bail, B., Munteanu, M., ... & Bourlière, M. (2007). Variability of the area under the receiver operating characteristic curves in the diagnostic evaluation of liver fibrosis markers: impact of biopsy length and fragmentation. *Alimentary Pharmacology & Therapeutics, 25*(6), 733–739.

[13] Rose, N. C., Kaimal, A. J., Dugoff, L., Norton, M. E., & American College of Obstetricians and Gynecologists. (2020). Screening for fetal chromosomal abnormalities: ACOG practice bulletin, number 226. *Obstetrics & Gynecology, 136*(4), e48–e69.

[14] Cabello-Solorzano, K., Ortigosa de Araujo, I., Peña, M.,

Correia, L., & J. Tallón-Ballesteros, A. (2023, August). The impact of data normalization on the accuracy of machine learning algorithms: a comparative analysis. In *International conference on soft computing models in industrial and environmental applications* (pp. 344-353). Cham: Springer Nature Switzerland. [CrossRef]

[15] Sarker, M. (2024). Revolutionizing healthcare: the role of machine learning in the health sector. *Journal of Artificial Intelligence General science (JAIGS)* ISSN: 3006-4023, 2(1), 36-61. [CrossRef]

[16] Zhang, Y., Zheng, Y., Wang, D., Gu, X., Zyphur, M. J., Xiao, L., ... & Deng, Y. (2025). Shedding Light on the Black Box: Integrating Prediction Models and Explainability Using Explainable Machine Learning. *Organizational Research Methods*, 10944281251323248. [CrossRef]

[17] Sadegh-Zadeh, S. A., Nazari, M. J., Aljamaeen, M., Yazdani, F. S., Mousavi, S. Y., & Vahabi, Z. (2024). Predictive models for Alzheimer's disease diagnosis and MCI identification: The use of cognitive scores and artificial intelligence algorithms. N*PG Neurologie-Psychiatrie-Gériatrie, 24*(142), 194-211. [CrossRef]

[18] Javaid, M., Haleem, A., Singh, R. P., Suman, R., & Rab, S. (2022). Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks, 3*, 58-73. [CrossRef]

[19] Sadegh-Zadeh, S. A., Sakha, H., Movahedi, S., Harandi, A. F., Ghaffari, S., Javanshir, E., ... & Hajizadeh, R. (2023). Advancing prognostic precision in pulmonary embolism: a clinical and laboratory-based artificial intelligence approach for enhanced early mortality risk stratification. *Computers in Biology and Medicine, 167*, 107696. [CrossRef]

[20] Sadegh-Zadeh, S. A., Soleimani Mamalo, A., Kavianpour, K., Atashbar, H., Heidari, E., Hajizadeh, R., ... & Gargari, S. S. (2024). Artificial intelligence approaches for tinnitus diagnosis: leveraging high-frequency audiometry data for enhanced clinical predictions. *Frontiers in Artificial Intelligence, 7*, 1381455. [CrossRef]

[21] Steffensen, E. H., Pedersen, L. H., Lou, S., Vogel, I., Danish Fetal Medicine Study Group, & Danish Cytogenetic Central Registry Study Group. (2023). Is the first-trimester combined screening result associated with the phenotype of Down syndrome? A population-based cohort study. *Prenatal Diagnosis, 43*(1), 51-61. [CrossRef]

[22] Valentini, D., Di Camillo, C., Mirante, N., Vallogini, G., Olivini, N., Baban, A., ... & Villani, A. (2021). Medical conditions of children and young people with Down syndrome. *Journal of Intellectual Disability Research, 65*(2), 199-209. [CrossRef]

[23] Wang, Q., Ma, Y., Zhao, K., & Tian, Y. (2022). A comprehensive survey of loss functions in machine learning. *Annals of Data Science, 9*(2), 187-212. [CrossRef]

[24] Yang, Y., & Xu, Z. (2020). Rethinking the value of labels for improving class-imbalanced learning. *Advances in Neural Information Processing Systems, 33*, 19290–19301. [CrossRef]

[25] Zhang, J., Ma, X., Zhang, J., Sun, D., Zhou, X., Mi, C., & Wen, H. (2023). Insights into geospatial heterogeneity of landslide susceptibility based on the SHAP-XGBoost model. *Journal of environmental management, 332*, 117357. [CrossRef]