STUDYING CLASS IMBALANCE FOR PRETERM BIRTH PREDICTION FROM THE COMPUTERIZED ANALYSIS OF TRANSVAGINAL ULTRASOUND IMAGES

NATALIA JOHANA CABEZA GUTIÉRREZ

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍAS ELÉCTRICA, ELECTRÓNICA Y DE TELECOMUNICACIONES
BUCARAMANGA
2026

STUDYING CLASS IMBALANCE FOR PRETERM BIRTH PREDICTION FROM THE COMPUTERIZED ANALYSIS OF TRANSVAGINAL ULTRASOUND IMAGES

NATALIA JOHANA CABEZA GUTIÉRREZ

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Electronic Engineering

Advisor
SAID DAVID PERTUZ ARROYO
ELECTRONIC ENGINEER. PhD

Co-advisor
CARLOS AUGUSTO FAJARDO ARIZA
ELECTRONIC ENGINEER. PhD

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍAS ELÉCTRICA, ELECTRÓNICA Y DE TELECOMUNICACIONES
BUCARAMANGA
2026

DEDICATION

To my family, professors, and to those who accompanied me with love and patience throughout this journey. Thank you for your constant support.

ACKNOWLEDGMENTS

I would like to thank my parents, who have always been the inspiration that allows me to pursue my goals. My mom, Hersilia, has supported me through every stage of my academic journey, understanding me in ways no one else can and giving me strength during both good and difficult moments. My dad, Víctor, even without fully understanding what I studied, always believed in me and offered his unconditional support. I am also deeply grateful to my siblings —Víctor Manuel, Arbelay, Deisy, and Sebastián— for being present throughout this path. I would also like to thank Jeison for his understanding, patience, and love during the most demanding moments. His constant encouragement made this journey much more manageable. Finally, I want to express my sincere gratitude to my advisor for his guidance, support, and commitment throughout the development of this thesis.

Thanks to all these wonderful people—family, friends, and mentors—for their love, support, and belief in me.

CONTENTS

Pag.

LIST OF FIGURES

Pag.

LIST OF TABLES

Pag.

LIST OF ANNEXES

Pag.

RESUMEN

TÍTULO ESTUDIANDO EL DESBALANCE DE CLASES PARA LA PREDICCIÓN DEL PARTO PREMATURO A PARTIR DEL ANÁLISIS COMPUTARIZADO DE IMÁGENES DE ULTRASONIDO TRANSVAGINAL [*]

AUTOR: Natalia Johana Cabeza Gutiérrez [**]

PALABRAS CLAVE: Parto prematuro, ultrasonido transvaginal, redes neuronales convolutivas, desbalance de clases, primer trimestre.

DESCRIPCIÓN: El parto prematuro (PP) continúa siendo una de las principales causas de mortalidad y morbilidad neonatal. La detección temprana es fundamental para identificar a las mujeres en riesgo y reducir las complicaciones asociadas. Aunque existen diversos biomarcadores para estimar el riesgo de PP, la mayoría se centra en el segundo trimestre, cuando los cambios cervicales son más evidentes. Además, los algoritmos basados en aprendizaje automático presentan limitaciones debido al marcado desbalance de clases, lo que genera modelos sesgados, con baja sensibilidad y altas tasas de falsos negativos. Con el fin de abordar esta limitación, en este estudio se propone un algoritmo para la predicción de PP a partir de imágenes de ultrasonido transvaginal del primer trimestre, incorporando estrategias existentes en el estado del arte para la compensación de desbalance de clases. Para ello, se implementó una red neuronal convolucional (RNC) basada en EfficientNetB4 con estrategias de compensación, empleada en una cohorte retrospectiva de 253 mujeres. La combinación de la arquitectura RNC con borderline-SMOTE obtuvo un AUC de 0.701 (IC: 0.552–0.825), mientras que la función focal loss logró un mejor equilibrio entre sensibilidad y especificidad. Estos hallazgos confirman que abordar el desbalance de clases aumenta la capacidad discriminativa y la sensibilidad del modelo. Asimismo, se evaluó el valor predictivo de la RNC junto con variables clínicas, y todas las configuraciones mostraron mejoras, destacándose la entropía cruzada ponderada con el mejor AUC (0.784; IC: 0.683–0.868). Hasta donde se conoce, este es el primer análisis con aprendizaje profundo aplicado a imágenes transvaginales del primer trimestre para estimar el riesgo de PP, destacando el potencial de biomarcadores tempranos basados en imágenes.

---

[*] Tesis

[**] Facultad de Ingenierías Fisicomecánicas. Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones. Director: Said Pertuz, PhD. Codirector: Carlos Fajardo, PhD.

ABSTRACT

TITLE: STUDYING CLASS IMBALANCE FOR PRETERM BIRTH PREDICTION FROM THE COMPUTERIZED ANALYSIS OF TRANSVAGINAL ULTRASOUND IMAGES [*]

AUTHOR: Natalia Johana Cabeza Gutierrez [**]

Keywords: Preterm birth, transvaginal ultrasound, convolutional neural networks, class imbalance, first trimester.

Description: Preterm birth (PTB) remains one of the leading causes of neonatal mortality and morbidity. Early detection is crucial to identify women at risk and thereby reduce the incidence and complications caused by PTB. Although there are different biomarkers that attempt to identify the risk of PTB, most focus on the second trimester when structural and biophysical changes in the cervix are most noticeable. Moreover, existing machine learning based-algorithms are limited by severe class imbalance, which leads to models biased toward the majority class (non-PTB), with low sensitivity and high false negative rates. To address this limitation, in this study we propose an algorithm for PTB prediction from first-trimester transvaginal ultrasound images that incorporates state-of-the-art compensation strategies to mitigate the effects of class imbalance. For this, a convolutional neural network (CNN) based on the EfficientNetB4 architecture with class imbalance compensation was implemented in a retrospective cohort of 253 women. The obtained results show that using class imbalance compensation strategies improves PTB prediction. The CNN architecture combined with the borderline-SMOTE strategy obtained an AUC of 0.701 (CI: 0.552, 0.825), whereas using the focal loss function resulted in a better trade-off between sensitivity and specificity. These findings confirm that addressing class imbalance enhances the model's performance by improving its discrimination ability and sensitivity. Additionally, the predictive value of the CNN combined with clinical variables was tested, and all strategies showed improved performance, with the weighted cross-entropy function obtaining the best AUC with a value of 0.784 (CI: 0.683, 0.868). To our knowledge, this is the first deep learning-based analysis of first-trimester TVUS images for PTB risk assessment, highlighting the potential of early imaging-based biomarkers for preventive obstetric care.

---

[*] Thesis

[**] Facultad de Ingenierías Fisicomecánicas. Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones. Advisor: Said Pertuz, PhD. Co-advisor: Carlos Fajardo , PhD.

INTRODUCTION

Births that occur before 37 weeks of gestation are considered preterm birth (PTB). PTB can be medically induced because of health conditions of the mother, such as preeclampsia and diabetes, or spontaneous (sPTB) due to premature rupture of membranes and cervical dilatation; this latter accounting for about 70 % of premature births[1].

Globally, 15 million children are born premature each year [2], and despite current advances in medicine, PTB is still the leading cause of neonatal death in children under 5 years, since about one million babies die due to PTB complications each year [3]. Moreover, children who survive to PTB are also exposed to other adverse effects, including infections, respiratory and digestive complications that also affect their quality of life[4].

According to the World Health Organization (WHO), no region of the world has significantly reduced the incidence rates of PTB over the last decade, since between 2010 and 2020 the reduction was just 0.14 % [3]. The frequency of PTB varies worldwide between 9 % to 12 %[2]. As stated in a report by the Departamento Administrativo Nacional de Estadística, 11.1 % of live births in Colombia during 2023 were preterm. Santander is one of the departments with higher rates of prematurity, since about 113 births out of 1000 were

---

[1] T. Włodarczyk et al.: *Machine Learning Methods for Preterm Birth Prediction: A Review*. En: *Electronics* 10 (2021). DOI: 10.3390/electronics10050586.

[2] C. P. Howson et al.: *Born too soon: preterm birth matters*. En: *Reproductive Health* (2013). DOI: 10.1186/1742-4755-10-S1-S1

[3] World Health Organization: *Preterm Birth*. 2023. https://www.who.int/news-room/fact-sheets/detail/preterm-birth

[4] H. Blencowe et al.: *National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications*. En: *The Lancet* (2012). DOI: 10.1016/S0140-6736(12)60820-4.

preterm[5].

In order to reduce health problems and life loss caused by PTB, early detection becomes crucial because it allows timely treatment [6], including progesterone administration and cervical cerclage. Transvaginal ultrasound (TVUS) is the gold-standard screening technique for pregnant women since it allows a detailed evaluation of the cervix, a region that exhibits structural and biophysical changes that can be associated with the risk of preterm delivery[7]. For this reason, different biomarkers obtained during the TVUS evaluation have emerged as indicators for PTB risk estimation, including cervical length (CL)[8], widely used in clinical practice, and cervical consistency index (CCI)[9]. However, these measurements rely exclusively on the experience of the physician, which might result in intra and inter-reader variability[10]. Besides, it has been shown that CL has low sensitivity, detecting between 50 % and 60 % of PTBs in women at low risk[11]. Thus, there is the need for

---

5 Departamento Administrativo Nacional de Estadística: *Estadísticas Vitales (EEVV) - Nacimientos en Colombia*. 2024. https://www.dane.gov.co/files/operaciones/EEVV/bol-EEVV-Nacimientos-IVtrim2023.pdf.

6 T. Włodarczyk et al.: *Spontaneous Preterm Birth Prediction Using Convolutional Neural Networks*. En: *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis* (2020)

7 R. A. Word et al.: *Dynamics of cervical remodeling during pregnancy and parturition: mechanisms and current concepts*. En: *Seminars in reproductive medicine* (2007). DOI: 10.1055/s-2006-956777.

8 J.D. Lams et al.: *The length of the cervix and the risk of spontaneous premature delivery. National Institute of Child Health and Human Development Maternal Fetal Medicine Unit Network*. En: *The New England Journal of Medicine* 334 (1996). DOI: 10.1056/NEJM199602293340904..

9 M. Parra-Saavedra et al.: *Prediction of preterm birth using the cervical consistency index.* En: *Ultrasound Obstet Gynecol* 38 (2011), págs. 44-51. DOI: 10.1002/uog.9010.

10 L. Valentin e I. Bergelin: *Intra- and interobserver reproducibility of ultrasound measurements of cervical length and width in the second and third trimesters of pregnancy*. En: *Ultrasound in Obstetrics and Gynecology* (2002). DOI: 10.1046/j.1469-0705.2002.00765.x.

11 F. L. Facco y H. N. Simhan: *Short ultrasonographic cervical length in women with low-risk obstetric history*. En: *Obstetrics and Gynecology* (2013). DOI: 10.1097/AOG.0b013e3182a2dccd.

more precise techniques that do not depend exclusively on the experience of the medical personnel.

With the rise of machine learning, different techniques have emerged such as classifiers obtained from risk factors, quantitative analysis of cervical texture and convolutional neural networks (CNNs). This latter has shown promising results in automating cervical texture analysis and improving PTB prediction, achieving area under the curve (AUC) values up to 0.75 [12]. However, existing approaches to date focus on the second trimester of pregnancy where cervical changes are more pronounced. This leaves a critical gap: no prior study has applied CNNs to first trimester TVUS for PTB risk assessment, despite the potential for earlier clinical decision-making. Only one related effort has explored image analysis in the first trimester using radiomic features [13], and none have leveraged deep learning for this purpose. The first trimester presents unique challenges —subtler anatomical changes and limited annotated data— but also a profound opportunity. Identifying risk at this stage could shift the clinical paradigm toward truly preventive care, reducing PTB incidence before symptoms manifest.

Moreover, current artificial intelligence based approaches are still limited by severe class imbalance, which is caused by the uneven distribution of classes in the datasets used for developing techniques for PTB prediction, since one out of ten births is preterm [14]. Training on such imbalanced datasets might result in suboptimal model performance for

---

[12] E. P. F. Sejer et al.: *The combined use of cervical ultrasound and deep learning improves the detection of patients at risk for spontaneous preterm delivery*. En: *American Journal of Obstetrics and Gynecology* (2025)

[13] W. Cancino, C. H. Becerra-Mojica y S. Pertuz: “Radiomic analysis of transvaginal ultrasound cervical images for prediction of preterm birth”. En: *Medical Image Understanding and Analysis (MIUA)*. Vol. 14860. Lecture Notes in Computer Science. 2024, págs. 414-424

[14] World Health Organization: *1 in 10 babies worldwide are born early, with major impacts on health and survival*. Accessed: Oct. 24, 2025. 2023. `https://www.who.int/news/item/06-10-2023-1-in-10-babies-worldwide-are-born-early--with-major-impacts-on-health-and-survival`

risk assessment since these could be biased towards the majority class and it might lead to the misclassification of the minority class, which in this case is the PTB class.

In other words, due to the inherent class imbalance in preterm birth datasets, risk assessment models may struggle to correctly identify preterm cases. This often results in a high false negative rate, thereby reducing the model's sensitivity and hindering early detection[15], which is critical for timely intervention.

In the literature, different techniques have been developed for addressing class imbalance in other classification tasks, and have yielded promising results [16][17]. These strategies range from algorithm-level approaches, which involve adjusting class weights or introducing loss functions that emphasize the minority class, to data-level techniques that balance the class distribution through undersampling of the majority class or oversampling of the minority class.

Motivated by the lack of studies on the prediction of preterm birth in the first trimester, particularly CNN-based, and the impact of existing class imbalance that affects the performance of the existing approaches, we address these gaps by designing an algorithm for PTB prediction from first-trimester TVUS images that incorporates state-of-the-art compensation strategies to mitigate the effects class imbalance in the PTB risk estimation context. Specifically, our contributions are twofold: (1) we pioneer deep learning-based analysis in an early gestational window that has been largely overlooked, and (2) we

---

[15] J. L. Leevy et al.: *A survey on addressing high-class imbalance in big data*. En: *Journal of Big Data* 5.1 (2018), pág. 42. DOI: 10.1186/s40537-018-0151-6.

[16] D. Dablain, B. Krawczyk y N. V. Chawla: *DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data*. En: *IEEE Transactions on Neural Networks and Learning Systems* 34 (2023), págs. 6390-6404. DOI: 10.1109/TNNLS.2021.3136503

[17] A. Witmer y B. Bhanu: *Iterative pseudo balancing for stem cell microscopy image classification*. En: *Scientific reports* 14 (2024). DOI: 10.1038/s41598-024-54993-y.

implement advanced class imbalance compensation strategies to enhance model sensitivity and reliability. In addition, we study the potential of combining our CNN-based analysis with clinical variables, including demographic information, cervical measurements and obstetric history, to PTB risk assessment.

1. BACKGROUND

Class imbalance is one of the main limitations in the construction of machine learning (ML) models. It is caused by the uneven distribution of classes in the datasets employed for training and validating these models, where one class —typically referred to as the majority class— outnumbers one or more minority classes. This limitation is particularly critical in medical applications, where data is inherently imbalanced due to the low incidence of the disease or outcome of interest. In the context of PTB, this imbalance arises naturally since approximately one in ten births is preterm[14].

When ML models are trained on such imbalanced distributions, they tend to be biased towards the majority class, non-PTB in our context. As consequence, their ability to correctly identify minority samples (PTB cases) decreases, leading to lower sensitivity and higher false negative rates. This issue is especially problematic for PTB prediction, where early detection is essential for timely interventions.

In the literature, different strategies have been proposed to mitigate the effects of class imbalance. These include algorithm-level approaches, such as assigning higher weights to the minority class or introducing new loss functions (e.g. focal loss[18]) during model training; and data-level techniques, such as oversampling the minority class using well-stablished methods like Synthetic Minority Oversampling Technique (SMOTE)[19] and its variant borderline-SMOTE[20]. Although these strategies have been widely used in medical

---

[18] T.-Y. Lin et al.: "Focal loss for dense object detection". En: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, págs. 2980-2988. DOI: 10.1109/ICCV.2017.324

[19] N. V. Chawla et al.: *SMOTE: Synthetic minority over-sampling technique*. En: *Journal of Artificial Intelligence Research* 16 (2002), págs. 321-357. DOI: 10.1613/jair.953

[20] H. Han, W. Y. Wang y B. H. Mao: "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning". En: *Advances in Intelligent Computing*. Vol. 3644. Lecture Notes in Computer Science.

contexts, their effectiveness may depend on the severity of class imbalance and the characteristics of each specific medical dataset. Therefore, understanding how class imbalance handling strategies perform under different scenarios is crucial for designing robust models in medical application domains.

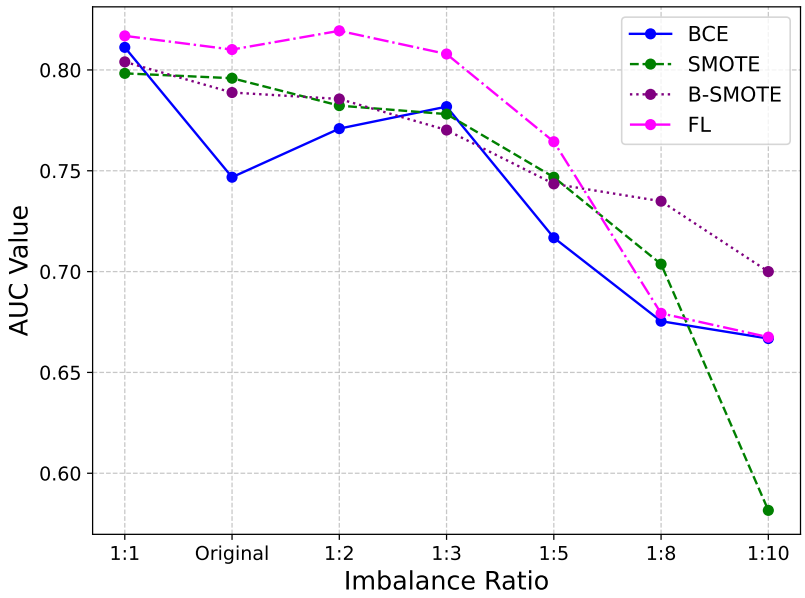
In order to study the behaviour of these techniques, we conducted a preliminary work in which we systematically evaluated strategies for compensating class imbalance in a medical classification task, specifically breast ultrasound lesion classification[21]. In this study, we started from a slightly imbalanced dataset (123 malignant vs 109 benign lesions) and generated six artificially imbalanced scenarios by applying random undersampling of the malignant class to simulate different disease incidences. We then evaluated the performance of focal loss, SMOTE and borderline-SMOTE across all the scenarios when these methods were applied to a logistic regression classifier trained using binary cross-entropy.

The obtained results exhibited that the model performance decreases as the imbalance ratio becomes more extreme, which was reflected by the decreasing trend of the AUC (see Fig. 1). In terms of sensitivity, we also observed that this metric decreased as the imbalance proportion increased, indicating that the model's ability to correctly identify malignant lesions worsened in scenarios with severe imbalance (see Fig. 2). When compensation strategies were incorporated, we observed performance improvements across all techniques, both in terms of AUC and sensitivity. These results confirmed that that the evaluated strategies are effective in mitigating the effects of class imbalance. However, our experiments also suggested that the selection of an appropriate class imbalance compensation strategy depends on the degree of imbalance. In subtler imbalance scenarios (1:2

Springer, 2005, págs. 878-887. DOI: 10.1007/11538059_91

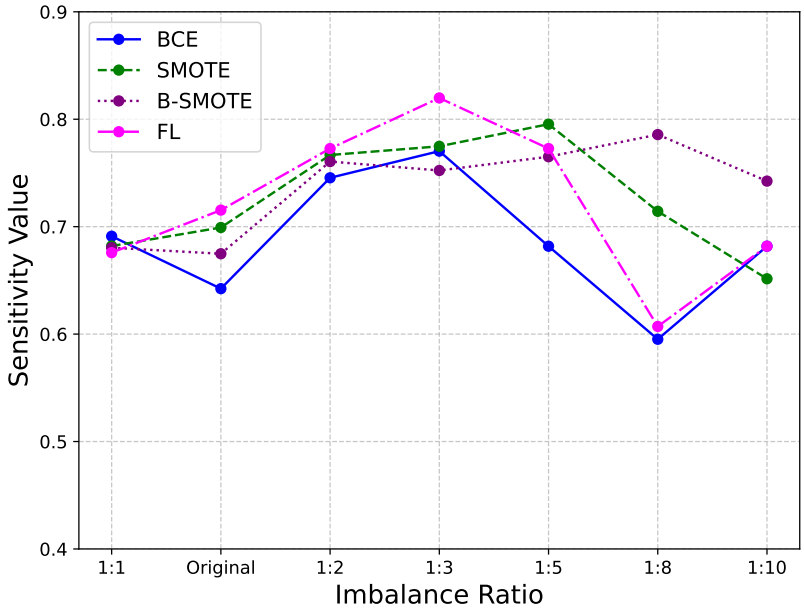
[21] N. Cabeza, C. A. Fajardo y S. Pertuz: "Evaluating the impact of class imbalance on breast ultrasound image classification". En: *Proc. XXV Symposium on Image, Signal Processing and Artificial Vision*. 2025, págs. 1-5

Figure 1. AUC scores for different imbalance ratios using all training configurations: BCE (Binary Cross-Entropy), SMOTE (Synthetic Minority Oversampling Technique), B-SMOTE (Borderline-SMOTE), and FL (Focal Loss).



to 1:5), focal loss achieved the best performance, meanwhile in more extreme scenarios bordeline-SMOTE outperformed both SMOTE and focal loss. These findings highlight that choice of compensation strategy should consider both the degree of imbalance, linked to the disease incidence, and the clinical relevance of the predicted outcome, the latter being particularly important when using oversampling techniques that modify the original data distribution.

Figure 2. Sensitivity scores for different imbalance ratios using all training configurations: BCE (Binary Cross-Entropy), SMOTE (Synthetic Minority Oversampling Technique), B-SMOTE (Borderline-SMOTE), and FL (Focal Loss).



2. OBJETIVES

2.1. GENERAL OBJECTIVE

To design an algorithm for preterm birth prediction from computerized analysis of transvaginal ultrasound (TVUS) images by incorporating strategies to address class imbalance.

2.2. SPECIFIC OBJECTIVES

1. To develop a baseline algorithm for preterm birth prediction from the computerized analysis of TVUS images.
2. To incorporate at least two state-of-the-art techniques for addressing class imbalance in the baseline model: one at the algorithmic level and one at the data level.
3. To evaluate and compare the performance of the incorporated techniques for addressing class imbalance in the construction of preterm birth risk assessment models from the computerized analysis of TVUS images.

3. METHODS

3.1. DATASET

The dataset for the development of this project is a retrospective cohort collected between November 2019 and July 2021 at two centers: Hospital Universitario de Santander and Centro Materno Fetal Inutero in Bucaramanga, Colombia. The study protocol and data collection procedures were approved by the Ethics Committee of the Universidad Industrial de Santander, and all participants provided informed consent prior to enrollment.

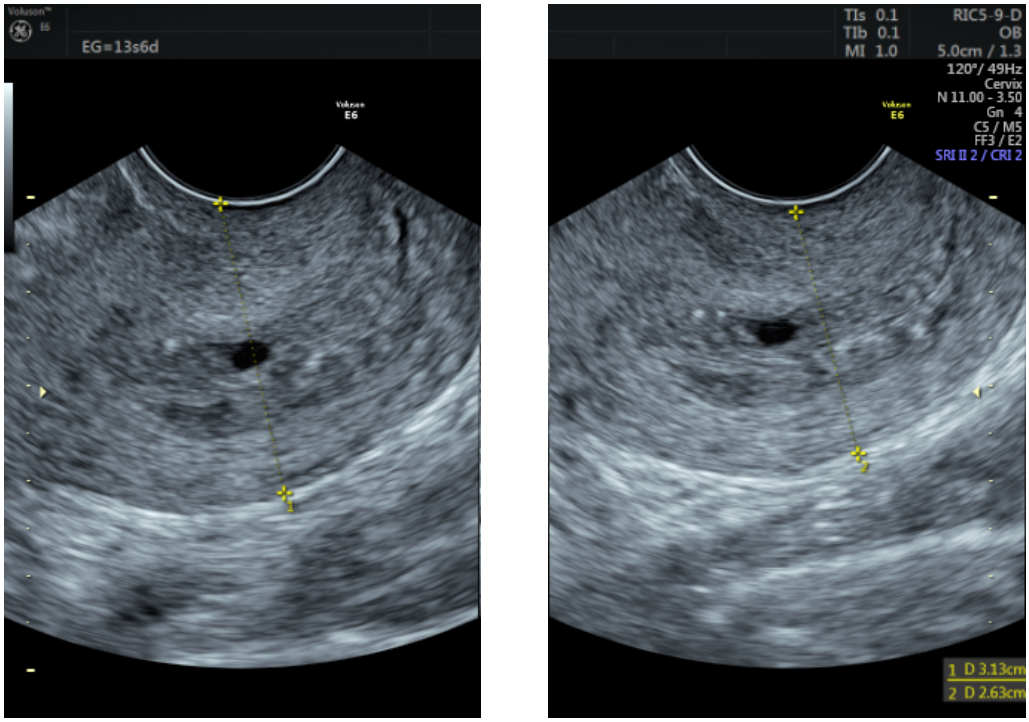
This dataset consists of first trimester TVUS images from 253 women, where 28 had spontaneous preterm birth and 225 had a full-term birth. A total of 504 images are available: 450 corresponding full-term births and 54 to spontaneous preterm births. Women included in this study were of majority age (over 18 years), had singleton pregnancies and spontaneous preterm deliveries. The acquisition of the TVUS images and cervical measures followed the protocol described by Becerra-Mojica et al,[22]. Each TVUS scan is composed of two images in sagittal view (see Fig. 3), one where the cervix is a rest or uncompressed, and other where the cervix is compressed by applying light pressure with the transducer. In both images, the anteroposterior diameters are measured to calculate the cervical consistency index (CCI). All the ultrasounds are accompanied by their respective outcome and clinical variables, including: age, height, weight, body mass index, history of previous PTB, number of previous PTB, cervical length (CL), anteroposterior diameters, and gestational age of cervical measurement.

---

[22] C. H. Becerra-Mojica et al.: *Cohort profile: Colombian Cohort for the Early Prediction of Preterm Birth (COLPRET): early prediction of preterm birth based on personal medical history, clinical characteristics, vaginal microbiome, biophysical characteristics of the cervix and maternal serum biochemical markers*. En: *BMJ Open* (2022). DOI: 10.1136/bmjopen-2021-060556.

In this work, we only used the uncompressed images for analysis, since these represent the cervix and preserve its natural morphology without transducer-induced deformation which is operator-dependent. However, for completeness, additional experimental results derived from the compressed images are presented in Annex. A.

Figure 3. TVUS image. The dash line correspond to the measurements of the cervix anteroposterior diameters, and the cross pointers are the callipers. (a) Uncompressed TVUS. (b) Compressed TVUS.



3.2. DATA PREPROCESSING

For the subsequent computerized analysis, data preprocessing is conducted to improve the quality of the TVUS images. This procedure includes two steps: first, pixel standardization since these images were acquired using different equipment and it might affect the analysis and comparison of these. Thus, all images were standardized to a pixel size of 0.21 mm to ensure high resolution and that no information is lost.

As a second step, artifact removal. This step involves detecting the field of view using morphological operations, which allow to detect the area of the tissue captured by the transducer, and elements such as the header or footer to be removed. Subsequently, unnecessary elements such as callipers, lines and text are removed. Eliminating these elements results in a loss of tissue information, therefore, the inpainting algorithm proporse by Anupam *et al.*[23] is used to generate representations of the missing tissue using surrounding pixels. Fig. 4 illustrates the result of this preprocessing pipeline on an uncompressed cervical image.

Figure 4. Example of preprocessed uncompressed TVUS image. The preprocessing pipeline includes pixel standardization and removal of annotation and artifacts, followed by inpainting to reconstruct missing tissue regions.

[23] A. Anupam, P. Goyal, and S. Diwakar: “Fast and enhanced algorithm for exemplar-based image inpainting”. In: *Proceedings of the 4th Pacific-Rim Symposium on Image and Video Technology (PSIVT)*. IEEE, 2010, pp. 325–330.

3.3. BASELINE MODEL ARCHITECTURE

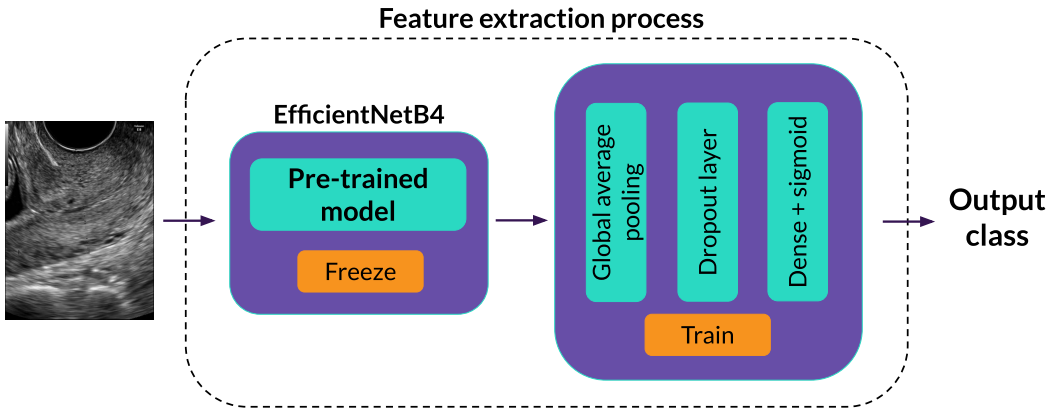
Given the small size of our sample (253 patients and 504 images), training a deep learning model from scratch in this case may not be a suitable option, as this model could have little predictive power and low generalization in our data. Therefore, we employed a transfer learning approach to take a model previously trained on other imaging contexts and adapt it to ours for the prediction of PTB. In this work, the transfer learning-based architecture and customized top layer proposed by Jha *et al.*[24], specifically the first step, was used for feature extraction and classification tasks. Based on experimental results (see Annex. B) EfficientNetB4 was implemented as backbone architecture, and was also initialized with pretrained weights from ImageNet.

The convolutional base of EfficientNetB4 was kept frozen to retain learned representations from ImageNet, preventing updates to its parameters during training. The output feature maps of this backbone were passed through a GlobalAveragePooling2D layer to reduce dimensionality while preserving spatial information. This was followed by a Dropout layer with a rate of 0.2, which helps to mitigate overfitting by randomly deactivating neurons during training. Finally, a fully connected dense layer with sigmoid activation function was added to produce the final output corresponding to the two target classes (preterm or normal birth). Fig 5 illustrates the proposed architecture with customized layer for preterm birth prediction.

---

[24] A. Jha, E. Jhon, and T. Banerjee: "Multi-class classification of dementia from MRI images using transfer learning". In: *Proceedings of the IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*. IEEE, 2022, pp. 1–6. DOI: 10.1109/UEMCON56746.2022.9997329.

Figure 5. Proposed model architecture. The CNN receives the preprocessed TVUS as input and feature extraction is conducted using EfficientNetB4 pre-trained network, followed by customized layers for performing preterm birth classification.



3.4. MODEL TRAINING AND HYPERPARAMETER OPTIMIZATION

The dataset was split at the patient level, with 70% of the patients used for training and validation, and the remaining 30% assigned to the test set. Then, data augmentation was applied only to the training images to improve generalization and enhance robustness to variations in the ultrasound data. Specifically, this data augmentation step included rotations of up to ±10°, translations in both height and width directions of up to 5%, and random contrast adjustments of up to 20%.

In order to find the most optimal model, a hyperparameter search was conducted on the training and validation subset using 5-fold cross-validation. A grid search strategy was implemented to systematically evaluate all combinations of learning rates [1e-3, 3e-3, 1e-4, 1e-5] and batch sizes [4, 8, 12, 16]. For each hyperparameter combination, the model was trained on four folds and validated on the remaining fold, rotating across folds. The highest mean validation AUC across the five folds was used as the selection criterion. The model was compiled with Adam optimizer and binary cross-entropy was used as loss function, with EarlyStopping (patience = 8) and ReduceLROnPlateau (min LR = 1e-6) to

avoid overfitting. The model with the best validation AUC was selected and evaluated on the independent test set.

3.5. CLASS IMBALANCE COMPENSATION

Given the existing 1:8 imbalance ratio in our dataset (28 PTB vs 225 non-PTB), we incorporated strategies to compensate for class imbalance, including both data and algorithm-level approaches.

For data-level strategies, we employed two oversampling strategies:

- **Synthetic minority oversampling technique (SMOTE):** It generates artificial samples of the minority class by interpolating between existing minority instances and their nearest neighbors in feature space. Specifically, for a selected minority sample, one or more synthetic samples are created along the line segment joining it to randomly chosen nearest neighbors. This approach increases the diversity of the minority class without simple duplication, helping to balance class distributions and improve classifier performance on imbalanced datasets[19].
- **Borderline-SMOTE (B-SMOTE):** It is a variation of the original SMOTE algorithm that concentrates on generating synthetic samples close to the decision boundary, particularly in regions where minority class instances are at higher risk of being misclassified. Instead of applying interpolation uniformly across all minority samples, B-SMOTE identifies instances whose nearest neighbors predominantly belong to the majority class and generates synthetic samples around them[20].

On the other hand, for algorithm-level strategies, we considered two loss functions commonly used to address class imbalance:

- **Weighted binary cross-entropy (WBCE):** It is a modification of the standard binary

cross-entropy that assigns different weights to each class in order to address class imbalance. A weighting factor w_t is applied to the loss term of each class t, increasing the contribution of underrepresented classes to the total loss[25].

$$\text{WBCE}(p_t) = -w_t \log(p_t) \tag{1}$$

where where p_t is the predicted probability associated with the true class. In this study, the minority class (PTB) was weighted according to the ratio between the number of majority and minority samples ($w_{\text{PTB}} = N_{\text{major}}/N_{\text{minor}}$), and the weights were normalized to maintain balanced loss contributions.

- **Focal loss function (FL):** It is is an extension of the standard binary cross-entropy designed to handle class imbalance by introducing two additional parameters: a balancing factor $\alpha \in [0, 1]$ and the focusing parameter $\gamma \geq 0$.

$$\text{FL}(p_t) = -\alpha(1 - p_t)^{\gamma} \log(p_t) \tag{2}$$

These parameters allow the loss to dynamically adjust the contribution of each sample based on its difficulty. When a sample is correctly classified (p_t close to 1), the modulating factor $(1 - p_t)^{\gamma}$ reduces its influence on the total loss. Conversely, for hard or misclassified samples (p_t low), the factor remains large, maintaining a higher loss value[18]. This mechanism effectively down-weights easy examples and focuses the learning process on challenging ones, improving the model's ability to learn from the minority class and enhancing performance under imbalanced conditions.

---

[25] Saining Xie and Zhuowen Tu: "Holistically-nested edge detection". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 1395–1403. DOI: 10.1109/ICCV.2015.164.

3.6. PREDICTION WITH CLINICAL VARIABLES

To evaluate the added predictive value of the CNN-derived imaging features over clinical variables, a hybrid model combining TVUS–based information with clinical variables (CV) was implemented. After training the CNN described in Section 3.3, the CNN-derived scores were extracted for all patients and concatenated with their corresponding clinical variables, which included age, weight, body mass index (BMI), history of previous PTB, number of previous PTBs, gestational age at cervical measurement, CL, CCI, and anteroposterior diameters. This combined feature set was then used to train a support vector machine (SVM) classifier, with hyperparameters optimized through grid search on the training and validation sets.

3.7. PERFORMANCE METRICS

To evaluate the model's performance, three metrics typically used in PTB risk assessment were calculated: AUC, sensitivity, and specificity. For the sensitivity and specificity metrics, these values were computed at a false positive rate (FPR) of 15%. In addition, the adjusted odds ratio (OR*) was computed to quantify the contribution of the CNN-derived scores to the hybrid models that combine clinical variables with image-based features.

4. RESULTS

This section summarizes the experimental results obtained from the proposed CNN-based model for PTB prediction using first-trimester TVUS images. Initially, the performance of the baseline model employing uncompressed TVUS images is reported, followed by the results obtained when applying class imbalance compensation strategies such as SMOTE, Borderline-SMOTE, weighted binary cross-entropy (WBCE), and focal loss (FL) in Table 1. Our propose baseline model achieved an AUC value of 0.612 (CI: 0.451, 0.758), sensitivity and specificity of 0.25 and 0.86, respectively. All strategies improved the model performance by increasing all performance metrics, however, borderline-SMOTE provided the best AUC with a value of 0.701 (CI: 0.552, 0.825). While borderline-SMOTE demonstrated notable discriminatory capacity and all compensation methods increased sensitivity, FL function achieved the highest sensitivity (0.56), effectively doubling the value obtained by the baseline model.

Table 1. Performance comparison of the baseline convolutional neural network (CNN) and class imbalance compensation strategies for preterm birth prediction. FL: Focal Loss, WBCE: Weighted Binary Cross-Entropy, SMOTE: Synthetic Minority Over-sampling Technique, B-SMOTE: Borderline-SMOTE. Values correspond to AUC (95% CI), sensitivity and specificity at FPR=15%.

Imbalance compensation	AUC	Sensitivity	Specificity
Baseline	0.612 [0.451, 0.758]	0.25	0.86
Algorithm-based			
FL	0.659 [0.492, 0.809]	**0.56**	0.85
WBCE	0.615 [0.430, 0.774	0.44	**0.87**
Data-based			
SMOTE	0.697 [0.545, 0.817]	0.44	0.84
B-SMOTE	**0.701 [0.552, 0.825]**	0.44	0.85

Following these findings, the predictive capability of our CNN-based model was further assessed within a hybrid approach that integrated clinical variables to explore the added

Table 2. Performance comparison of the proposed CNN combined with clinical variables (CV) and different class imbalance compensation strategies for preterm birth (PTB) prediction using uncompressed TVUS images. FL: Focal Loss, WBCE: Weighted Binary Cross-Entropy, SMOTE: Synthetic Minority Over-sampling Technique, B-SMOTE: Borderline-SMOTE. Values correspond to AUC, sensitivity and specificity (FPR = 15%), and adjusted odds ratio (OR*) with their respective 95% CI.

Scenario	AUC	Sensitivity	Specificity	OR*
CV	0.671 [0.534, 0.784]	0.50	0.86	-
Algorithm-based				
FL + CV	0.759 [0.643, 0.856]	0.38	0.86	1.498 [0.803, 2.795]
WBCE + CV	**0.784 [0.683, 0.868]**	0.38	0.85	1.517 [0.820, 2.808]
Data-based				
SMOTE + CV	0.720 [0.599, 0.828]	0.38	0.86	1.703 [0.897, 3.235]
B-SMOTE + CV	0.719 [0.595, 0.828]	0.38	0.86	1.720 [0.903, 3.274]

predictive value of combining imaging and clinical information. The corresponding results are summarized in Table 2. All hybrid configurations outperformed the model using only clinical variables, as evidenced by their higher AUC values and adjusted odds ratios (OR* > 1), demonstrating the predictive contribution of the CNN-based features. Among them, the best overall performance was achieved by the hybrid model incorporating WBCE as class imbalance compensation strategy, reaching an AUC of 0.784 (CI: 0.683, 0.868). However, the model with clinical variables alone exhibited the highest sensitivity.

5. DISCUSSIONS

5.1. MAIN FINDINGS

Our work shows the potential of CNN-based computerized ultrasound image analysis for the prediction of preterm birth. Our baseline model achieved an AUC value of 0.612 (CI: 0.451, 0.758), which is attributed to the existing class imbalance in our dataset caused by PTB incidence. However, when class imbalance compensation strategies are incorporated, the model's performance improves in terms of its discriminatory ability (AUC) and there is a decrease in the false negative rate (better sensitivity). Among the evaluated strategies, borderline-SMOTE achieved the best performance (AUC = 0.701 (CI: 0.552, 0.825)). However, when using the focal loss function, better sensitivity was obtained: increasing from 0.25 in our baseline model to 0.56.

These findings demonstrate that strategies such as SMOTE, borderline-SMOTE, focal loss and weighted binary cross-entropy are effective in mitigating the effects of class imbalance by improving the model's predictive power when only using an image-based approach. However, the selection of these depends on the specific clinical objective of the task and the incidence of the outcome of interest [21]. In the context of PTB risk estimation, improving the model's discriminative ability (reflected by higher AUC values) may not always coincide with enhancing its sensitivity, which is crucial from a clinical perspective since failing to correctly identify women at risk of PTB could have significant implications.

5.2. PREDICTIVE VALUE WITH CLINICAL VARIABLES

Our hybrid model that combines CNN-based predictions with clinical variables demonstrated a clear improvement in the predictive performance for PTB risk assessment compared to using clinical variables alone. This improvement is reflected in higher AUC values

across all evaluated scenarios, where the model with WBCE and including clinical variables achieved the best performance (AUC = 0.784 (CI:0.683, 0.868)). Although all hybrid configurations yielded adjusted odds ratios greater than 1 (OR* $>$1), their wide confidence intervals indicate that these effects were not statistically significant. Therefore, we cannot conclude that the association between the CNN-derived imaging features and PTB risk is independent of the clinical variables.

Furthermore, sensitivity did not improve when incorporating the CNN with clinical variables. It might be partially attributed to the limited size of our test set: with only a few positive PTB cases available, where even a single misclassification could lead to a large proportional drop. It suggests that the lack of improvement in sensitivity does not necessarily imply a lack of predictive benefit from the CNN-based predictor, but rather reflects the constraints imposed by the small number of positive cases.

5.3. COMPARISON WITH OTHER WORKS

Previous works have used computerized techniques, particularly CNN based approaches, to analyze cervical texture and predict PTB. One study employed the U-Net network for the prediction of preterm birth by only using ultrasound scans and achieved an AUC of 0.72[6]. Another study combined DTU-Net and SonoNet for PTB risk assessment and achieved an AUC of 0.75 [12], also relying exclusively on image-based information. Although these results are promising, these works were developed and evaluated on second-trimester ultrasound scans, when structural changes in the cervix are more pronounced. In contrast, our work is focused on the first trimester, an earlier gestational window in which cervical anatomical differences are subtler.

To date, only one prior study has explored first-trimester PTB risk assessment from TVUS images. Cancino *et al.* proposed a radiomics-based analysis that combines image features with clinical variables to improve prediction [13]. In contrast, our work assesses PTB

risk by employing a CNN-based model with transfer learning to address the limitations imposed by a small sample size, and to our knowledge this is one of the first works that address PTB prediction in the first trimester employing deep learning-based approaches.

5.4. STRENGTHS AND LIMITATIONS

As mentioned above, one of the main strengths of our work is that we perform preterm birth risk assessment in the first trimester, a gestational window that has been scarcely explored, but holds high clinical value for enabling earlier detection and timely intervention. Cervical changes are subtler during this period, making PTB prediction more challenging, and our study is among the first to apply deep learning–based techniques for this task.

Another contribution of our work lies in addressing the class imbalance inherent to PTB epidemiology. We evaluated four compensation strategies and demonstrated their effectiveness in mitigating the adverse effects of class imbalance, as reflected in the improved performance of our baseline model. When using only image-based information, our results achieved values comparable to previously published state-of-the-art approaches. Additionally, we assessed the predictive value of our CNN-derived scores and showed that they provide meaningful added value to clinical variables (OR* $>$ 1).

However, our study also has certain limitations. First, our sample contains a limited number of PTB cases (28 in total), and after splitting into training and test sets, this latter subset contained only eight positive cases. This scarcity of PTB samples introduces statistical uncertainty, which is reflected in the width of the confidence intervals across the evaluated metrics. Under this scenario, misclassifying a single PTB sample has a marked effect on the performance metrics. For example, in our test set (8 PTBs), one incorrect prediction reduces the sensitivity by 12.5 percentage points, and two misclassifications lead to a drop of 25 percentage points. This arguably explains why our hybrid models did not exhibit improvements in sensitivity, even though they were superior in AUC values.

Although the adjusted odds ratios were consistently greater than 1, indicating a potential added predictive value of the CNN-derived imaging features, their 95% confidence intervals included 1. Statistically, this implies that the null hypothesis of no added predictive value ($H_0 : OR^* = 1$) cannot be rejected. Thus, despite our results suggesting a beneficial effect, the evidence is not strong enough to conclude that this improvement is statistically significant. This lack of significance is a consequence of the small number of positive cases, which reduces the precision of the estimated effect size and leads to wide confidence intervals. Furthermore, such limited data availability also has methodological implications in terms of model generalization. The reduced number of PTB cases increases the risk of overfitting, which is inherent to modeling low-prevalence outcomes such as preterm birth. With so few PTB samples, the model might learn patterns that are too specific to these particular cases rather than capturing generalizable imaging features.

We implemented strategies to mitigate the risk of overfitting and enhance the model's ability to generalize. Data splitting was performed at the patient level to prevent information leakage between training and test sets, and model performance was assessed on an independent hold-out test. Additionally, cross-validation within the training set, data augmentation and callbacks, including early stopping, were employed to improve model robustness. However, although the dataset comprised information from two clinical centers, the evaluation remains limited to this specific cohort, and no external validation was conducted. Therefore, future work should include validation on independent cohorts and diverse acquisition environments to further assess the generalizability of the proposed framework.

Beyond these limitations, additional methodological strategies could be further strengthened in our proposed framework. First, although we incorporated well-established class imbalance compensation strategies, we did not include more advanced approaches such as DeepSMOTE [16]. In this study, we focused on widely adopted state-of-the-art class

imbalance mitigation techniques, motivated both by their extensive use in the literature and by the findings of our preliminary study [21], as our objective was to provide a first structured approach to addressing class imbalance in the context of first trimester preterm birth prediction using deep learning. Additionally, considering the limited sample size of our cohort, employing imbalance mitigation methods of moderate complexity, such as SMOTE and Borderline-SMOTE, represented a more practical and stable approach for this setting. Nevertheless, future research should explore the integration of more advanced strategies given that recent literature has highlighted their potential to improve the representation of minority classes in medical tasks.

Finally, we did not incorporate a model interpretability analysis, such as Gradient-weighted Class Activation Mapping (Grad-CAM)[26], to visualize the image regions that contribute the most to the CNN predictions. The inclusion of such explanaibility techniques would allow for a better understanding of whether the model is focusing on meaningful cervical regions, which is important for clinical reliability of this approach. Incorporating these tools represents an important direction for future work, particularly in medical applications where transparency and interpretability are essential for clinical adoption.

---

[26] R. Selvaraju et al.: "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74.

6. CONCLUSIONS

In this work, we proposed a CNN-based model with transfer learning and class imbalance compensation strategies for preterm birth prediction from first-trimester TVUS images of the cervix. This early-stage approach shows the potential of deep learning to extract informative image-based features before clinical symptoms appear. Given the intrinsic class imbalance in our dataset, we incorporated compensation strategies to improve model performance. Among the evaluated strategies, borderline-SMOTE achieved the highest AUC (0.701 [CI: 0.552, 0.825]), whereas focal loss provided a better trade-off between sensitivity and specificity, an essential aspect in clinical applications where missing a case has significant implications.

To the best of our knowledge, this is the first deep learning-based approach for PTB risk assessment using first-trimester cervical images. These findings demonstrate that addressing class imbalance enhances the reliability of first-trimester only image-based PTB risk assessment, highlighting the potential of early ultrasound-based biomarkers for preventive obstetric care. Furthermore, by evaluating the incremental predictive value of the CNN-derived scores when combined with clinical variables, we showed that our imaging-based model contributes complementary discriminatory information beyond traditional clinical predictors.

The full source code used for model development, training, and evaluation is publicly available in the following GitHub repository: `https://github.com/nataliacabeza/Studying-Class-Imbalance-for-PTB-Prediction.git`

CONTRIBUTIONS

Conference proceedings

- Cabeza, N., Fajardo C. A., & Pertuz, S. (2025). Evaluating the Impact of Class Imbalance on Breast Ultrasound Image Classification. 2025 XXV Symposium of Image, Signal Processing, and Artificial Vision (STSIVA), pp. 1-5. DOI: 10.1109/STSIVA66383.2025.11156656.
  Status: Published.

- Cabeza, N., Fajardo, C. A., Becerra-Mojica, C.H., & Pertuz, S. CNN-Based Prediction of Preterm Birth from First-Trimester Transvaginal Ultrasound. IEEE International Symposium on Biomedical Imaging (ISBI 2026).
  Status: Under review.

Collaborations

- Valenzuela, M. F., Cabeza, N., & Pertuz, S. Does research on ai for medical imaging adhere to reporting quality standards?. IEEE International Symposium on Biomedical Imaging (ISBI 2026).
  Status: Under review.

Bibliography

Anupam, A., P. Goyal, and S. Diwakar: “Fast and enhanced algorithm for exemplar-based image inpainting”. In: *Proceedings of the 4th Pacific-Rim Symposium on Image and Video Technology (PSIVT)*. IEEE, 2010, pp. 325–330.

Becerra-Mojica, C. H. et al.: *Cohort profile: Colombian Cohort for the Early Prediction of Preterm Birth (COLPRET): early prediction of preterm birth based on personal medical history, clinical characteristics, vaginal microbiome, biophysical characteristics of the cervix and maternal serum biochemical markers*. In: *BMJ Open* (2022). DOI: 10.1136/bmjopen-2021-060556.

Blencowe, H. et al.: *National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications*. In: *The Lancet* (2012). DOI: 10.1016/S0140-6736(12)60820-4.

Cabeza, N., C. A. Fajardo, and S. Pertuz: “Evaluating the impact of class imbalance on breast ultrasound image classification”. In: *Proc. XXV Symposium on Image, Signal Processing and Artificial Vision*. 2025, pp. 1–5.

Cancino, W., C. H. Becerra-Mojica, and S. Pertuz: “Radiomic analysis of transvaginal ultrasound cervical images for prediction of preterm birth”. In: *Medical Image Understanding and Analysis (MIUA)*. Vol. 14860. Lecture Notes in Computer Science. 2024, pp. 414–424.

Chawla, N. V. et al.: *SMOTE: Synthetic minority over-sampling technique*. In: *Journal of Artificial Intelligence Research* 16 (2002), pp. 321–357. DOI: 10.1613/jair.953.

Dablain, D., B. Krawczyk, and N. V. Chawla: *DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data*. In: *IEEE Transactions on Neural Networks and Learning Systems* 34 (2023), pp. 6390–6404. DOI: 10.1109/TNNLS.2021.3136503.

Departamento Administrativo Nacional de Estadística: *Estadísticas Vitales (EEVV) - Nacimientos en Colombia*. 2024. https://www.dane.gov.co/files/operaciones/EEVV/bol-EEVV-Nacimientos-IVtrim2023.pdf.

Facco, F. L. and H. N. Simhan: *Short ultrasonographic cervical length in women with low-risk obstetric history*. In: *Obstetrics and Gynecology* (2013). DOI: 10.1097/AOG.0b013e3182a2dccd.

Han, H., W. Y. Wang, and B. H. Mao: "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning". In: *Advances in Intelligent Computing*. Vol. 3644. Lecture Notes in Computer Science. Springer, 2005, pp. 878–887. DOI: 10.1007/11538059_91.

Howson, C. P. et al.: *Born too soon: preterm birth matters*. In: *Reproductive Health* (2013). DOI: 10.1186/1742-4755-10-S1-S1.

Jha, A., E. Jhon, and T. Banerjee: "Multi-class classification of dementia from MRI images using transfer learning". In: *Proceedings of the IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*. IEEE, 2022, pp. 1–6. DOI: 10.1109/UEMCON56746.2022.9997329.

Lams, J.D. et al.: *The length of the cervix and the risk of spontaneous premature delivery. National Institute of Child Health and Human Development Maternal Fetal Medicine Unit Network*. In: *The New England Journal of Medicine* 334 (1996). DOI: 10.1056/NEJM199602293340904..

Leevy, J. L. et al.: *A survey on addressing high-class imbalance in big data*. In: *Journal of Big Data* 5.1 (2018), p. 42. DOI: 10.1186/s40537-018-0151-6.

Lin, T.-Y. et al.: "Focal loss for dense object detection". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 2980–2988. DOI: 10.1109/ICCV.2017.324.

Parra-Saavedra, M. et al.: *Prediction of preterm birth using the cervical consistency index.* In: *Ultrasound Obstet Gynecol* 38 (2011), pp. 44–51. DOI: 10.1002/uog.9010.

Sejer, E. P. F. et al.: *The combined use of cervical ultrasound and deep learning improves the detection of patients at risk for spontaneous preterm delivery*. In: *American Journal of Obstetrics and Gynecology* (2025).

Selvaraju, R. et al.: "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74.

Valentin, L. and I. Bergelin: *Intra- and interobserver reproducibility of ultrasound measurements of cervical length and width in the second and third trimesters of pregnancy*. In: *Ultrasound in Obstetrics and Gynecology* (2002). DOI: 10.1046/j.1469-0705.2002.00765.x.

Witmer, A. and B. Bhanu: *Iterative pseudo balancing for stem cell microscopy image classification*. In: *Scientific reports* 14 (2024). DOI: 10.1038/s41598-024-54993-y.

Włodarczyk, T. et al.: *Machine Learning Methods for Preterm Birth Prediction: A Review*. In: *Electronics* 10 (2021). DOI: 10.3390/electronics10050586.

Włodarczyk, T. et al.: *Spontaneous Preterm Birth Prediction Using Convolutional Neural Networks*. In: *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis* (2020).

Word, R. A. et al.: *Dynamics of cervical remodeling during pregnancy and parturition: mechanisms and current concepts*. In: *Seminars in reproductive medicine* (2007). DOI: 10.1055/s-2006-956777.

World Health Organization: *1 in 10 babies worldwide are born early, with major impacts on health and survival*. Accessed: Oct. 24, 2025. 2023. https://www.who.int/news/item/06-10-2023-1-in-10-babies-worldwide-are-born-early--with-major-impacts-on-health-and-survival.

World Health Organization: *Preterm Birth*. 2023. https://www.who.int/news-room/fact-sheets/detail/preterm-birth.

Xie, Saining and Zhuowen Tu: "Holistically-nested edge detection". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 1395–1403. DOI: 10.1109/ICCV.2015.164.

ANNEXES

Annex A. Results for compressed TVUS images

In addition to the experiments conducted on uncompressed TVUS images, which were selected as the primary analysis due to the absence of operator-induced tissue deformation, this annex reports the corresponding results obtained when evaluating the baseline model and the class imbalance compensation strategies on compressed TVUS images. Table 3 summarizes the obtained results for all evaluated configurations.

Table 3. Performance comparison of the baseline convolutional neural network (CNN) and class imbalance compensation strategies for preterm birth prediction in compressed images. FL: Focal Loss, WBCE: Weighted Binary Cross-Entropy, SMOTE: Synthetic Minority Over-sampling Technique, B-SMOTE: Borderline-SMOTE. Values correspond to AUC (95% CI), sensitivity and specificity at FPR=15%.

Imbalance compensation	**AUC**	**Sensitivity**	**Specificity**
Baseline	0.652 [0.514, 0.775]	0.19	0.85
Algorithm-based			
FL	0.660 [0.537, 0.772]	0.19	0.85
WBCE	0.657 [0.527, 0.782]	0.31	0.85
Data-based			
SMOTE	0.744 [0.624, 0.847]	0.38	0.85
B-SMOTE	0.742 [0.622, 0.846]	0.38	0.84

Despite obtaining good results when using SMOTE (AUC = 0.744 [0.624, 0.847]) and borderline-SMOTE (AUC = 0.742 [0.622, 0.846]), it is evident that all strategies obtain lower sensitivity compared to the results obtained when uncompressed images were used. Because of this and due to the deformation induced by the operator-dependent transducer,

we opted to use images where there was no compression of the cervix and its natural morphology was preserved.

Annex B. Evaluation of different backbone architectures for preterm birth prediction

To determine the most suitable backbone architecture for the proposed deep learning model, we conducted a systematic comparison of state-of-the-art convolutional neural network backbones: EfficientNetB4, DenseNet169, Xception and InceptionV3. This evaluation aimed to identify the architecture that provided the best balance between predictive performance, computational efficiency, and stability during training. Each candidate backbone was integrated into the same model pipeline, trained under identical conditions, and assessed using the same evaluation metrics to ensure a fair comparison. The results of these experiments, summarized in Table 4, guided the selection of the final backbone architecture used throughout this work.

Table 4. Performance comparison of candidate backbone architectures for preterm birth prediction using first-trimester TVUS images.

Architecture	**AUC**	**Sensitivity**	**Specificity**
EfficientNetB4	**0.612 [0.451, 0.758]**	**0.25**	0.86
InceptionV3	0.487 [0.355, 0.620]	0.06	0.85
Xception	0.516 [0.357, 0.680]	0.19	**0.87**
DenseNet169	0.375 [0.224, 0.533]	0.00	0.86

Based on these experimental results, we selected EfficientNetB4 as backbone architecture since it provided the higher AUC with a value of 0.612 (CI: 0.451, 0.7578) and better trade-off between sensitivity and specificity with results of 0.25 and 0.86, respectively.