

Ensemble of fine-tuned machine learning models for hysterectomy prediction in pregnant women using magnetic resonance images

Vishnu Vardhan Reddy Kanamata Reddy,^{a,b} Michael Villordon,^{a,b}
Quyên N. Do,^{b,c} Yin Xi,^{b,c} Matthew A. Lewis,^c Christina L. Herrera,^{d,e}
David Owen,^{b,d,e} Catherine Y. Spong,^{b,d,e} Diane M. Twickler,^{c,d,e}
and Baowei Fei^{b,a,b,c,*}

^aThe University of Texas at Dallas, Department of Bioengineering, Richardson, Texas, United States

^bThe University of Texas at Dallas, Center for Imaging and Surgical Innovation, Richardson, Texas, United States

^cThe University of Texas Southwestern Medical Center, Department of Radiology, Dallas, Texas, United States

^dThe University of Texas Southwestern Medical Center, Department of Obstetrics and Gynecology, Dallas, Texas, United States

^eParkland Health, Dallas, Texas, United States

ABSTRACT. Purpose: Identifying pregnant patients at high risk of hysterectomy before giving birth informs clinical management and improves outcomes. We aim to develop machine learning models to predict hysterectomy in pregnant women with placenta accreta spectrum (PAS).

Approach: We developed five machine learning models using information from magnetic resonance images and combined them with topographic maps and radiomic features to predict hysterectomy. The models were trained, optimized, and evaluated on data from 241 patients, in groups of 157, 24, and 60 for training, validation, and testing, respectively.

Results: We assessed the models individually as well as using an ensemble approach. When these models are combined, the ensembled model produced the best performance and achieved an area under the curve of 0.90, a sensitivity of 90.0%, and a specificity of 90.0% for predicting hysterectomy.

Conclusions: Various machine learning models were developed to predict hysterectomy in pregnant women with PAS, which may have potential clinical applications to help improve patient management.

© The Authors. Published by SPIE under a Creative Commons Attribution 4.0 International License. Distribution or reproduction of this work in whole or in part requires full attribution of the original publication, including its DOI. [DOI: [10.1117/1.JMI.12.2.024502](https://doi.org/10.1117/1.JMI.12.2.024502)]

Keywords: deep learning; radiomics; topographic feature maps; vision transformers; classification; magnetic resonance imaging; placenta accreta spectrum; hysterectomy

Paper 24219GRR received Jul. 17, 2024; revised Feb. 7, 2025; accepted Feb. 24, 2025; published Mar. 18, 2025.

1 Introduction

Placenta accreta spectrum (PAS) occurs when the placenta fails to separate from the uterus after delivery.^{1,2} PAS is affecting up to 1 in 272 pregnancies.³⁻⁶ In cases of PAS, there may be substantial bleeding, which requires immediate removal of the uterus to save the mother's life. These urgent operations result in significant blood loss and complications.⁷⁻⁹ Identification of patients

*Address all correspondence to Baowei Fei, bfei@utdallas.edu

at risk of hysterectomy improves outcomes.^{10,11} Given the rising incidence rate of cesarean delivery and PAS, this is clinically significant.¹²

Convolutional neural networks (CNNs) have performed admirably in the field of computer vision.¹³ Numerous methods based on 2D and 3D CNN have been developed, particularly for the analysis of 3D medical images.^{14–17} 3D-based techniques can naturally learn 3D representations.^{18–20} In the meantime, transformer networks are frequently utilized in both computer vision^{21–23} and natural language processing.^{24,25} More recently, the pure transformer-based computer vision architecture, Vision Transformer (ViT),^{22,26} was able to reach state-of-the-art performance on image classification tasks. CNNs have significant inductive biases and locals that enable them to perform well even with little data. However, these biases may limit their effectiveness when dealing with high-dimensional data that require coverage beyond their typical low receptive field.^{23,27} The layout of a transformer architecture, with its modest inductive biases, can cover a vast region with a high receptive field. On the other hand, these biases might pose limitations when working with small datasets.^{22,28,29}

Recently, research has been done on a hybrid network, which combines CNN- and transformer-based architectures, to benefit from both approaches and produce more competitive performances in comparison to traditional approaches.^{27–29} In the field of medical imaging, there are fewer datasets available than in other fields due to ethical concerns,^{30,31} high computational costs,³² expensive annotation,³² and significant class-imbalance issues.³³ However, recent studies showed that combining 3D CNNs, 2D CNNs, and transformers leads to synergic effects, with the resulting network achieving state-of-the-art results for 3D medical image classification even with a smaller dataset.³⁴

Radiomics, a novel machine learning approach, endeavors to quantify phenotypic traits on medical images through automated algorithms, facilitating the extraction of a high-dimensional set of characteristics from clinical data for quantitative analysis of radiologic data.^{35,36} In earlier research, it was discovered that placenta and uterus volumes segmented using deep learning and expert segmentation shared roughly 40% of the same radiomic characteristics.^{37–39} Using a combination of radiomics and deep learning, a different group was able to predict placental invasion on T2-weighted magnetic resonance imaging (MRI) with 0.94 accuracy.⁴⁰ In a study employing radiomics to predict hysterectomy due to PAS, the best model had an area under the curve (AUC) of 0.80 for a cohort of 62 individuals.^{41,42} A study employing MRI to predict hysterectomy and PAS in pregnant women achieved a classification accuracy of 0.92 and 0.88, respectively, using radiomic characteristics derived from the placenta and uterine in 241 pregnant women's magnetic resonance (MR) images.⁴³ Deep learning methods using the different modalities of data, such as 3D MRI, 2D topographic feature maps, and radiomics achieved promising results in predicting PAS in pregnant women severe enough to warrant a hysterectomy after infant delivery.⁴⁴ Architectures, such as CascadeNet, were used to process the multimodal data for hysterectomy prediction and achieved an AUC of 0.878, accuracy of 83.3%, sensitivity of 85.0%, and specificity of 82.5%.⁴⁴

Building upon the promising results of combining deep learning and radiomics for hysterectomy prediction, this study proposes a novel multilevel ensemble approach. Through the use of ensemble learning, data from various classification models are combined into a single, superior classifier to produce improved prediction performance.⁴⁵ Expert-level predictions are provided for challenging medical image classification through the ensemble of neural networks.^{46,47} We leveraged the strengths of 3D CNNs, transformers, and statistical machine learning models by processing 3D MRI scans, topographic feature maps, and radiomics through five individual models. By combining the predictions from these models using a hard voting mechanism, we achieved superior performance in predicting the need for hysterectomy in pregnant patients with prenatal concern for PAS. This approach has the potential to improve clinical decision-making and patient outcomes by providing expert-level predictions.

2 Methods

2.1 MRI Data

The data used in our study consist of 241 T2-weighted MRI data (1.5T) from 241 pregnant women, grouped by their clinical outcome: those who eventually had a hysterectomy (88) and

those who did not (153). The size of the axial images was 256×256 pixels, with the exception of three patients where the in-plane sizes were zero-padded to 256×256 pixels. The slice thickness was 7.0 mm, with the total number of slices per patient ranging from 28 to 62. The in-plane resolution of the axial images ranged from $1.055 \times 1.055 \text{ mm}^2$ to $1.953 \times 1.953 \text{ mm}^2$ across all patients. The number of slices varies because the size of the uteruses varies among different subjects. Following our previous work by Dormer et al.,⁴⁴ who used the same dataset, the through-plane variability was addressed during preprocessing using an isotropic voxel size. This approach has been shown to maintain acceptable quality while enabling consistent 3D analysis across varying fields of view.

The 241 patients were split into groups of $N = 157$ for training, $N = 24$ for validation, and $N = 60$ for testing. The training data consist of 56 hysterectomy cases and 101 no-hysterectomy cases. The validation data consist of 12 hysterectomy cases and 12 no-hysterectomy cases, and finally, the test data consist of 20 hysterectomy cases and 40 no-hysterectomy cases.

2.2 Preprocessing

The 3D MRI volumes were resized to $192 \times 192 \times 25$ voxels using linear interpolation. Sample 3D MRI slices of the placenta are shown in Fig. 1. The training dataset had an imbalanced number of patients, so an additional technique was used to augment the data. This involved either cropping or padding the original MRI in the in-plane direction by 15% before resizing using linear interpolation to $192 \times 192 \times 25$ voxels. All 56 hysterectomy patients in the training dataset were augmented using both methods, resulting in a total of 168 volumes. For 101 no-hysterectomy cases, those with resolutions below $1.3 \times 1.3 \text{ mm}^2$ were augmented using the cropping method, whereas those with higher resolutions were augmented using the padding method. This doubled the number of normal patients in the training group to 202 and brought the total training dataset to 370 patients. This improved the ratio of hysterectomy to normal patients from roughly 1:2 to 4:5.⁴⁴

2.3 Radiomic Features

A radiologist's manual segmentations of the uterus and placenta were obtained and used as two individual masks to extract the radiomic features. In total, 214 radiomic features were extracted using PyRadiomics, including shape, gray level co-occurrence, gray level run length, and first-order statistics. A total of 107 features were extracted from each of the placenta and uterus regions.⁴⁴

2.4 Topographic Feature Map

To gain information about the placenta's texture and surrounding areas, novel topographic feature maps were made.⁴⁸ We used the distance between surface points and point of view as the displayed feature. Figure 2 shows a sample topographic map. The same topography-based scanning and mapping method was employed to extract and map various features, such as distance,

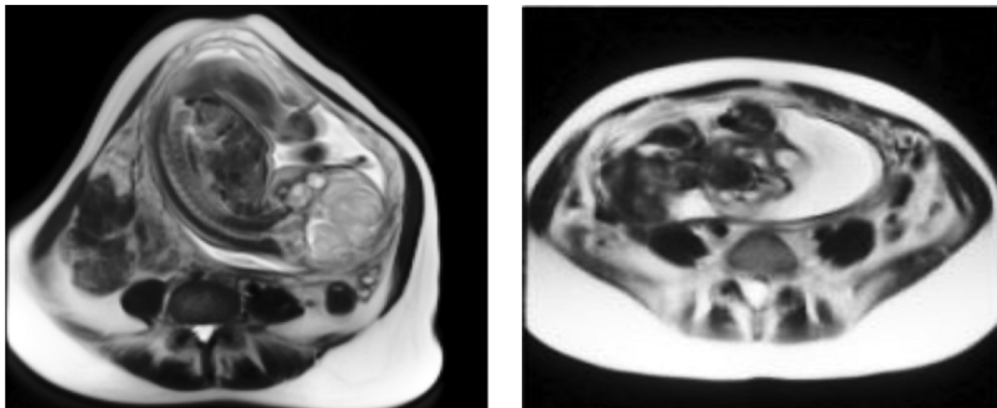


Fig. 1 3D MRI slices of the placenta.

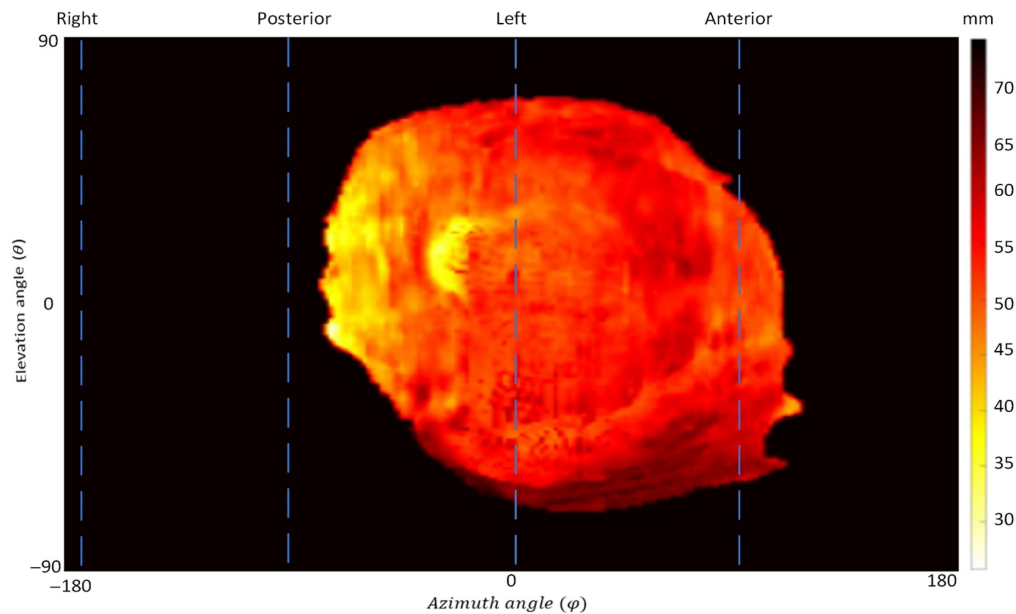


Fig. 2 Topographic feature map of the placenta. The superior and inferior sides are depicted, respectively, in the top and bottom halves of the topography map.

placenta thickness, surface intensity, local mean of intensities, and local standard deviation of intensities.⁴⁸ The patch size for the patch-based extraction was $11 \times 11 \times 11$ voxels.

2.5 Major Configurations of the Models

To comprehensively assess the potential of machine learning for predicting hysterectomy in pregnant patients, we developed and compared five distinct model configurations. The first method utilized traditional statistical machine learning on radiomic features extracted from MRI data, establishing a baseline performance benchmark. Next, we fine-tuned the CascadeNet⁴⁴ model with additional improvements. In a previous study,⁴⁴ it was mentioned that their method can be improved with additional fine tuning. For the third method, we explored the potential of transformer-based architectures by replacing the 2D CNN component of CascadeNet with a ViT for processing topographic feature maps derived from MRI data. To address potential data biases and enhance generalizability, we implemented stratified cross-validation on both fine-tuned CascadeNet model and CascadeNet model with ViT. This exploration of various machine learning approaches and data modalities allowed for a systematic evaluation of their effectiveness and culminated in the development of an ensemble model for predicting the need for a hysterectomy.

2.5.1 Model 1—machine learning classifier based on radiomics

In this method, we performed the classification task using 15 machine learning algorithms. The scikit-learn⁴⁹ package in Python was used to implement all machine learning methods with default settings. In addition, we used a machine learning classifier (LightGBMClassifier⁵⁰) implemented utilizing its respective Python package with its default parameters. For this method, we utilized radiomics features extracted from a combined dataset of 394 MR images (370 images from 157 patients in the training set and 24 images from 24 patients in the validation set) to fit all 16 classifiers. There were no patients in the training and testing groups who were the same. Their predictive performance was assessed on the testing set through accuracy, sensitivity, specificity, and ROC-AUC analysis.

After extensive training, we chose the LightGBM classifier as model 1 because it was the only machine learning classifier with high accuracy and AUC. Training only on radiomics allowed the model to have a different understanding of the data than the other ones, allowing for a unique generalization compared with all the other models.

2.5.2 Model 2—modified CascadeNet

The modified CascadeNet architecture is illustrated in Fig. 3, which shows three parallel processing paths: (1) the MRI path utilizing 3D convolutions for volumetric analysis of 3D MRI data, (2) the radiomics path consisting of dense layers for processing radiomic features, and (3) the topographic feature map path employing 2D convolutions, detailed in Fig. 4. Our key innovation lies in the feature fusion mechanism: we introduced additional deep neural network (DNN) layers after the concatenation of features from all three paths. These novel DNN layers were specifically designed to learn complex interactions among different modalities (MRI, radiomics, and topographic features) and extract higher-order patterns from their combined representations, thereby enhancing the model’s classification capabilities. This multimodal feature integration approach represents a significant advancement over the original CascadeNet architecture.⁴⁴

The DNN consists of a multilayer perceptron (MLP) with three layers each with 2000 features followed by the SeLU⁵¹ activation layer and Lecun normal kernel initialization⁵¹ followed

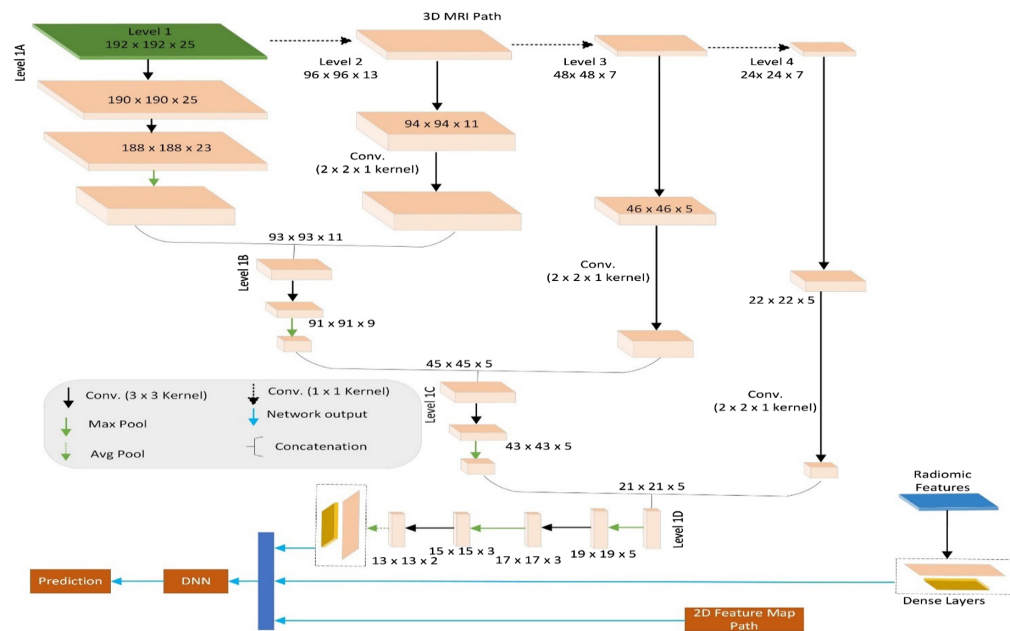


Fig. 3 Illustration of modified three-path version of CascadeNet.

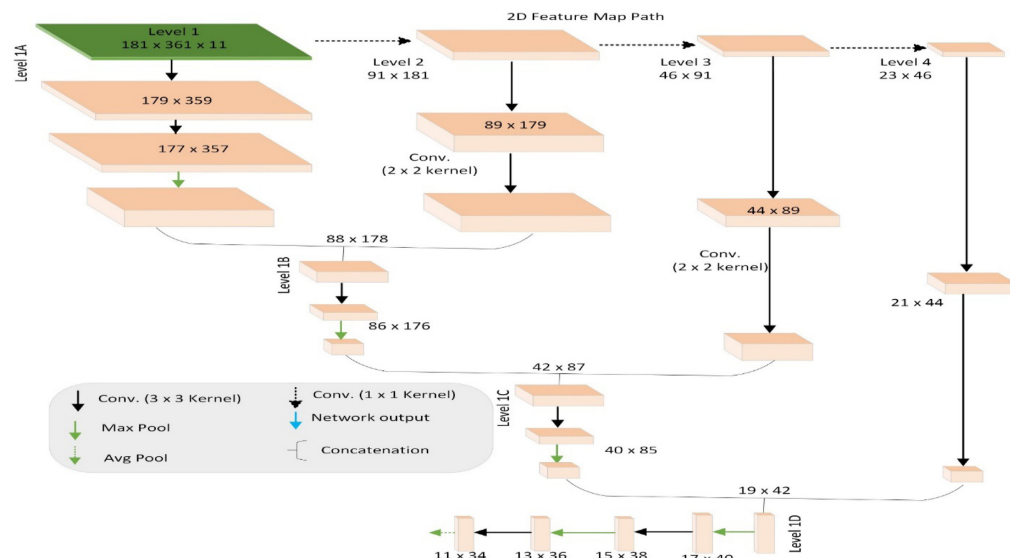


Fig. 4 Illustration of 2D topographic feature map processing component of CascadeNet.

by alpha dropout⁵¹ of 0.25. Finally, the prediction layer was a dense layer with two output features followed by a softmax activation function to classify the multimodal input. This modified CascadeNet improved the accuracy of test data by 2% compared with the previous CascadeNet⁴⁴ architecture.

2.5.3 Model 3—modified CascadeNet using cross-validation

To enhance model generalization and address potential data biases, we implemented an alternative training strategy using a stratified sevenfold cross-validation technique⁵² for model selection. In this approach, we combined our training and validation datasets and split them into seven folds while maintaining the class distribution in each fold. The modified CascadeNet was then trained seven times. This process yielded seven different models; each was evaluated on its respective validation fold. The model that achieved the highest performance on its validation fold was selected as our final model and was subsequently evaluated on the independent test set.

2.5.4 Model 4—replacing the 2D CNN component of the CascadeNet with ViT

We also developed an alternative version of our modified CascadeNet where we integrated ViT architecture specifically for processing the 2D topographic feature maps. In this configuration, although the 3D MRI path (with 3D convolutions) and the radiomics path remain unchanged as shown in Fig. 3, we replaced the 2D CNN component with a ViT architecture. The output features from the ViT are concatenated with the features from the MRI and radiomics paths and then processed through our novel DNN layers before final classification. This architectural variation was explored with two objectives: (1) to investigate whether transformer-based feature extraction could capture different patterns in the topographic features compared with traditional 2D CNNs and (2) to evaluate if the self-attention mechanism of ViT could provide complementary information when integrated with features from other modalities. However, this replacement with the ViT network only achieved comparable performance with the modified CascadeNet.

The ViT consists of four transformer layers with a patch size of 6, a hidden size of 128 units, and two heads. The MLP block in the ViT consists of two dense layers of 128 units each followed by a dropout layer of 0.15 and GeLU⁵³ activation. The final logit layer consisted of a dense layer of 150 units with ReLU⁵⁴ activation, and the weights were initialized to 0.

2.5.5 Model 5—replacing 2D CNN component of CascadeNet with ViT with cross-validation

The next model we evaluated consisted of the same architecture as model 4 but with stratified cross-validation. We used this technique for model selection by combining both training and validation datasets and training the model for seven folds. As we obtained seven estimates of the model's performance on its respective validation data, we chose the model with the highest performance out of seven estimates.

2.6 Implementation

Each network was trained using the RMSProp optimizer in TensorFlow, with an initial learning rate of 0.00001, adjusted through exponential decay. Empirical results demonstrated that RMSProp achieved faster convergence compared with other optimizers, including Adam and SGD. In addition, RMSProp's adaptive learning rate was advantageous given the data characteristics, aligning with findings reported by Dormer et al.⁴⁴ Binary cross entropy was used as the loss function, with a weighting factor of 1:1.15 (no hysterectomy: hysterectomy) to account for the imbalanced classes. During training, the dataset is shuffled after every epoch. Due to the nonsymmetric and complex nature of the data, as also noted by Dormer et al.,⁴⁴ who used the same dataset, additional augmentation technique was not employed. The networks were built on a CentOS 7 system with TensorFlow version 2.4 running in Docker. The training was conducted on an NVIDIA A6000 GPU. The evaluation metric used was patient-level accuracy. Once the ideal model for each method was found, these models were combined and evaluated on the reserved testing dataset.

2.7 Evaluation Metrics

We use accuracy, sensitivity, and specificity to evaluate the performance of the prediction model. Accuracy is defined as

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}, \quad (1)$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative. Specificity is the proportion of true negatives that the model correctly predicts, whereas sensitivity is the fraction of true positives that the model correctly predicts.

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})}, \quad (2)$$

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})}. \quad (3)$$

The area under the receiver operating characteristic curve using the probabilistic prediction was also used to evaluate the overall performance of the prediction model.

3 Results

3.1 Comparison Study Results

The results from radiomics test data when evaluated with 16 machine learning models are shown in Table 1. In our evaluation, we found that the LightGBM classifier outperformed every other model in terms of accuracy and ROC-AUC. Out of the 16 machine learning models, there were 10 models that had at least 70% accuracy and ROC-AUC.

Table 1 Quantitative analysis of ML models using radiomic features.

Network architecture	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
LightGBM	83.3	60.0	95.0	77.5
LogisticRegression	70.0	70.0	70.0	70.0
LogisticRegressionCV	71.6	70	72.5	71.2
PassiveAggressiveClassifier	45.0	55.0	40.0	47.5
Perceptron	66.6	0.0	100.0	50.0
KNeighborsClassifier	46.6	55.0	42.5	48.7
SVC	71.6	55.0	80.0	67.5
MLPClassifier	56.6	50.0	60.0	55.0
DecisionTreeClassifier	73.3	70.0	75.0	72.5
XGBClassifier	76.6	55.0	87.5	71.2
AdaBoostClassifier	73.3	60.0	80.0	70.0
SGDClassifier	66.6	0.0	100.0	50.0
RandomForestClassifier	75.0	55.0	85.0	70.0
GradientBoostingClassifier	76.6	65.0	82.5	73.7
ExtraTreeClassifier	61.6	60.0	62.5	61.2
RidgeClassifier	71.6	70	72.5	71.2

AUC, area under the curve; SVC, support vector classifier; MLP, multilayer perceptron; LGBM, light gradient boosting machine; SGD, stochastic gradient descent.

Table 2 Testing results of the three-path CascadeNet method.

Network architecture	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
CascadeNet	83.3	85.0	82.5	87.8
Model 1—LightGBM	83.3	60.0	95.0	77.5
Model 2—modified CascadeNet	85.0	90.0	82.5	86.2
Model 3—modified CascadeNet with CV	88.3	85.0	90.0	87.5
Model 4—modified CascadeNet with ViT	83.3	80.0	85.0	82.5
Model 5—modified CascadeNet with CV and ViT	85.0	70.0	92.5	81.2
Ensemble (average)	90.0	90.0	90.0	90.0

We chose the LightGBM model and the four modified three-path CascadeNet models and compared each model with the CascadeNet.⁴⁴ The quantitative performance is presented in Table 2. Our developed models outperformed the CascadeNet model in 50% of testing (bolded). We have also shown that the CascadeNet can be improved with additional architectural changes (model 3), fine-tuning, and cross-validation. However, when combining all five model's binary predictions through a majority voting scheme, the ensemble outperformed all its individual models' performances, giving state-of-the-art results.

In addition, the transformer-based architectures, such as ViT, had lower performance metrics than the modified CascadeNet model, which is of pure CNNs when trained on a small medical image dataset. Although each network achieved comparable performance with CascadeNet,⁴⁴ the sensitivity and specificity metrics show that these networks were able to uncover patterns that were able to confirm the hysterectomy in the event of a positive result. Combining these five models (ensemble) boosted metrics such as accuracy and AUC, thereby achieving robust performance.

The LightGBM classifier (model 1), which was trained only on radiomics data, has achieved high specificity, low sensitivity, and comparable accuracy when compared with other models. This statistical model may underestimate the need for a hysterectomy, potentially failing to identify patients who actually require the procedure.

To evaluate the performance improvements of our proposed models over the baseline three-path CascadeNet, we conducted statistical testing, with p -values provided in Table 3. Our ensemble model showed notable improvement. Although the p -value does not meet the conventional threshold for statistical significance (e.g., $p < 0.05$), the ensemble outperformed individual model variations. This finding suggests that the dataset of 60 samples may be too small to detect significant differences. Our individual models did not exhibit differences in performance, and the ensemble method leveraged complementary model strengths, indicating a potential for enhanced

Table 3 Statistical testing to evaluate the significance of the differences in performance metrics between three-path CascadeNet and each of our proposed models.

Network architecture	p -Value
Model 1—LightGBM	0.7518
Model 2—modified CascadeNet	1.0000
Model 3—modified CascadeNet with CV	0.2482
Model 4—modified CascadeNet with ViT	0.4795
Model 5—modified CascadeNet with CV and ViT	1.0000
Ensemble (average)	0.1336

Table 4 Impact of component models on ensemble performance in predicting hysterectomy with statistical comparison.

Ensemble	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	<i>p</i> -Value
Ensemble without model 1	85.0	70.0	92.5	81.2	0.3
Ensemble without model 2	86.7	80.0	90.0	85.0	0.4
Ensemble without model 3	85.0	70.0	92.5	81.2	0.3
Ensemble without model 4	88.3	95.0	85.0	90.0	1.0
Ensemble without model 5	86.7	80.0	90.0	85.0	0.4

robustness and generalization across varied test conditions. Although our proposed method achieved superior performance metrics (90% accuracy, AUC, sensitivity, and specificity compared with baseline), the statistical test suggests that more data might be needed to establish the statistical significance of these improvements.

3.2 Ablation Study

To evaluate the contribution of individual models within our proposed ensemble method, we conducted a comprehensive ablation study by systematically removing one model at a time from the ensemble while maintaining the averaging strategy. Table 4 presents the quantitative performance metrics and statistical analysis of each ablated version compared with the complete ensemble. The results demonstrate that removing model 1 or 3 had similar impacts on performance, with accuracy decreasing to 85.0% and AUC dropping to 81.2%. The removal of model 2 or 5 showed moderate performance changes with accuracy at 86.7% and AUC at 85.0%. Interestingly, the ensemble showed robust performance even with the removal of model 4, maintaining an accuracy of 88.3% and achieving an AUC of 90.0%. The *p*-values across all ablations suggest that although each model contributes to the ensemble's overall performance, the ensemble architecture maintains its robustness without critical dependence on any single model. These findings support our ensemble design choice, demonstrating that the combined approach effectively leverages the complementary strengths of different models while maintaining resilience to individual model variations.

4 Discussion and Conclusion

In this study, we developed five machine learning models based on MR images and combined them to predict hysterectomy in pregnant patients with increased risk of PAS using the topographic maps and radiomic features. The performance of the four versions of CascadeNet used on the multimodal data (MRI volumes, radiomic features, and custom feature maps) was significantly improved through architectural changes and fine-tuning, including vision transformers to extract meaningful insights from the custom feature maps. The statistical machine learning model (LightGBM classifier) using radiomics features alone achieved an accuracy of 83.3% (50/60 correct), with an AUC of 0.775. With the ensemble of the five models, we improved the AUC to 0.90 with an accuracy of 90.0%.

The patient populations involved were unbalanced, with the dataset of patients that eventually received a hysterectomy having fewer patients. This imbalance was addressed using techniques, such as stratified cross-validation and usage of class weights in loss function to improve model performance. Finally, we had done extensive testing and optimization to improve our networks, but due to the extremely limited size of our dataset, our test set only contained 60 patients. Our models would consistently get four cases/patients in our test set incorrect with high confidence in its predictions. As such, our models are incapable of correctly classifying those cases, and whatever they may represent, which shows that our training data are not diverse at this time. This could present problems for generalization. However, we have improved upon our previous work, increasing the accuracy by 7%, which shows that there is potential in our methodology and work.

Our approach incorporates the insights gained from Dormer et al.,⁴⁴ where adding topographic feature maps and radiomic features to 3D MRI data slightly enhanced model performance. Specifically, Dormer et al. demonstrated that the addition of each feature type yielded modest improvements in accuracy and AUC, with the highest metrics achieved when all features were combined. This prior research substantiates the predictive value of topographic and radiomic features, justifying their inclusion. Building upon their established feature combination framework, our work focuses on architectural innovations to better leverage these complementary features rather than re-validating their individual contributions. This approach allows us to concentrate on improving the model's ability to synthesize these proven feature sets more effectively.

Although the radiomic features are extracted from the images segmented by radiologists, which ensures precise identification of uterine and placental regions, this process can be directly applied to the images segmented by deep learning–based segmentation methods, as we previously published.^{38,39} Semi-automated or fully automated segmentation methods, leveraging recent advances in deep learning–based models, have shown promise in the image segmentation tasks and could be trained to identify uterine and placental regions with sufficient accuracy. Moreover, the deployment of automated methods for feature extraction may enhance scalability, allowing for more feasible integration into clinical workflows. Future work could focus on the development of such automated pipelines, which would facilitate the wider application of our predictive model in both large-scale research studies and clinical environments. As deep learning models for medical imaging continue to improve, we anticipate that automated segmentation could minimize radiologist intervention, making our approach more viable for routine clinical use.

We developed models 3 and 5 by employing stratified sevenfold cross-validation as a model selection technique rather than for performance estimation. This approach allowed us to train seven different versions of our modified CascadeNet architecture and modified CascadeNet architecture with ViT, each evaluated on its respective validation data. Although this methodology enabled us to select the model with the best generalization potential, we acknowledge certain limitations. As each fold's model was evaluated on different validation subsets, we cannot provide traditional cross-validation statistics such as performance standard deviations across folds. Our final performance metrics come from evaluating the selected best-performing model on a completely independent test set, which provides an unbiased assessment of model performance on unseen data. This approach prioritizes selecting the most robust model for clinical application, although it differs from traditional cross-validation approaches. Future work could explore ensemble methods or alternative validation strategies that might provide additional insights into model robustness while maintaining the benefits of our current approach.

Overall, our study highlights the potential of machine learning to predict hysterectomy in pregnant patients with prenatal concern for PAS. By combining models, the accuracy, sensitivity, specificity, and AUC improved. However, our work was not without limitations. Future work will explore using only 2D feature maps for training and prediction, expanding the patient population to improve the model's robustness. Combining convolutions and transformers with cross-attention mechanisms for processing all the modalities of data might improve the model's generalization. However, that would be a potential avenue for further exploration. Overall, this research demonstrates the potential of machine learning to pre-emptively predict hysterectomies during pregnancy in patients with PAS, which could have a life-saving impact and pave the way for further development in this field.

Disclosures

The authors have no relevant financial interests in this article and no potential conflicts of interest to disclose.

Code and Data Availability

Code and data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

Acknowledgments

Research reported in this publication was supported in part by the National Cancer Institute of the National Institutes of Health [Award Nos. R01CA288379 (BF), K25HD104004 (QND), K23HD103876 (CLH), and R01CA204254 (BF)] and the Cancer Prevention and Research Institute of Texas (CPRIT) [Award No. RP240289 (BF)]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. K. Benirschke and S. G. Driscoll, "Unusual shapes of the placenta: placenta accreta," in *The Pathology of the Human Placenta*, K. Benirschke and S. G. Driscoll, Eds., Springer, New York, pp. 106–134 (1967).
2. C. Maldjian et al., "MRI appearance of placenta percreta and placenta accreta," *Magn. Reson. Imaging* **17**(7), 965–971 (1999).
3. E. Jauniaux, J. C. Kingdom, and R. M. Silver, "A comparison of recent guidelines in the diagnosis and management of placenta accreta spectrum disorders," *Best Pract. Res. Clin. Obstet. Gynaecol.* **72**, 102–116 (2021).
4. S. Wu, M. Kocherginsky, and J. U. Hibbard, "Abnormal placentation: twenty-year analysis," *Am. J. Obstet. Gynecol.* **195**(5), 1458–1461 (2005).
5. M. F. Mogos et al., "Recent trends in placenta accreta in the United States and its impact on maternal-fetal morbidity and healthcare-associated costs," *J. Matern. Fetal Neonatal Med.* **29**(7), 1077–1082 (2016).
6. Publications Committee, Society for Maternal-Fetal Medicine, M. A. Belfort, "Placenta accreta," *Am. J. Obstet. Gynecol.* **203**(5), 430–439 (2010).
7. C. M. Briery et al., "Planned vs emergent cesarean hysterectomy," *Am. J. Obstet. Gynecol.* **197**(2), 154.e1–154.e5 (2007).
8. L. S. Machado, "Emergency peripartum hysterectomy: incidence, indications, risk factors and outcome," *N. Am. J. Med. Sci.* **3**(8), 358–361 (2011).
9. X. Kong et al., "On opportunity for emergency cesarean hysterectomy and pregnancy outcomes of patients with placenta accreta," *Medicine* **96**(39), e7930 (2017).
10. C. R. Warshak et al., "Effect of predelivery diagnosis in 99 consecutive cases of placenta accreta," *Obstet. Gynecol.* **115**(1), 65–69 (2010).
11. A. G. Eller et al., "Maternal morbidity in cases of placenta accreta managed by a multidisciplinary care team compared with standard obstetric care," *Obstet. Gynecol.* **117**(2 Pt 1), 331–337 (2011).
12. B. Liu et al., "Prediction of cesarean hysterectomy in placenta previa complicated with prior cesarean: a retrospective study," *BMC Preg. Childbirth* **20**(1), 81 (2020).
13. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM* **60**(6), 84–90 (2017).
14. T. Ni et al., "Elastic boundary projection for 3D medical image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, pp. 2109–2118 (2019).
15. H. R. Roth et al., "A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations," *Lect. Notes Comput. Sci.* **8673**, 520–527 (2014).
16. J. Yang et al., "Reinventing 2D convolutions for 3D images," *IEEE J. Biomed. Health Inform.* **25**(8), 3009–3018 (2021).
17. N. Bien et al., "Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet," *PLoS Med.* **15**(11), e1002699 (2018).
18. Ö. Çiçek et al., "3D U-Net: learning dense volumetric segmentation from sparse annotation," *Lect. Notes Comput. Sci.* **9901**, 424–432 (2016).
19. F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: fully convolutional neural networks for volumetric medical image segmentation," in *Fourth Int. Conf. 3D Vision (3DV)*, pp. 565–571 (2016).
20. H. R. Roth et al., "An application of cascaded 3D fully convolutional networks for medical image segmentation," *Comput. Med. Imaging Graph.* **66**, 90–99 (2018).
21. N. Carion et al., "End-to-end object detection with transformers," *Lect. Notes Comput. Sci.* **12346**, 213–229 (2020).
22. A. Dosovitskiy et al., "An image is worth 16 × 16 words: transformers for image recognition at scale," arXiv:2010.11929 (2020).
23. S. Khan et al., "Transformers in vision: a survey," *ACM Comput. Surv.* **54**(10s), 1–41 (2022).
24. A. Vaswani et al., "Attention is all you need," in *Adv. Neural Inf. Process. Syst.* **30**, pp. 6000–6010 (2017).
25. J. Devlin et al., "BERT: pre-training of deep bidirectional transformers for language understanding," in *North Am. Chapter Assoc. Comput. Linguist.*, pp. 4171–4186 (2019).
26. H. Touvron et al., "Training data-efficient image transformers & distillation through attention," in *Int. Conf. Mach. Learn.*, pp. 10347–10357 (2021).

27. Y. Zhao et al., “A battle of network structures: an empirical study of CNN, transformer, and MLP,” arXiv:2108.13002 (2021).
28. T. Xiao et al., “Early convolutions help transformers see better,” in *Adv. Neural Inf. Process. Syst.* **34**, pp. 30392–30400 (2021).
29. Z. Dai et al., “CoAtNet: marrying convolution and attention for all data sizes,” in *Adv. Neural Inf. Process. Syst.* **34**, pp. 3965–3977 (2021).
30. A. L. Simpson et al., “A large annotated medical image dataset for the development and evaluation of segmentation algorithms,” arXiv:1902.09063 (2019).
31. A. A. A. Setio et al., “Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge,” *Med. Image Anal.* **42**, 1–13 (2017).
32. N. Tajbakhsh et al., “Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation,” *Med. Image Anal.* **63**, 101693 (2020).
33. K. Yan et al., “Holistic and comprehensive annotation of clinically significant findings on diverse CT images: learning from radiology reports and label ontology,” in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, pp. 8523–8532 (2019).
34. J. Jang and D. Hwang, “M3T: three-dimensional medical image classifier using multi-plane and multi-slice transformer,” in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, pp. 20718–20729 (2022).
35. J. J. Van Griethuysen et al., “Computational radiomics system to decode the radiographic phenotype,” *Cancer Res.* **77**(21), e104–e107 (2017).
36. S. S. Yip and H. J. Aerts, “Applications and limitations of radiomics,” *Phys. Med. Biol.* **61**(13), R150 (2016).
37. Y. Xi et al., “Assessing reproducibility in magnetic resonance (MR) radiomics features between deep-learning segmented and expert manual segmented data and evaluating their diagnostic performance in pregnant women with suspected placenta accreta spectrum (PAS),” *Proc. SPIE* **11597**, 115972P (2021).
38. M. Shahedi et al., “Segmentation of uterus and placenta in MR images using a fully convolutional neural network,” *Proc. SPIE* **11314**, 113141R (2020).
39. M. Shahedi et al., “Deep learning-based segmentation of the placenta and uterus on MR images,” *J. Med. Imaging* **8**(5), 054001 (2021).
40. Q. Shao et al., “Deep learning and radiomics analysis for prediction of placenta invasion based on T2WI,” *Math. Biosci. Eng.* **18**(5), 6198–6215 (2021).
41. Q. N. Do et al., “Texture analysis of magnetic resonance images of the human placenta throughout gestation: a feasibility study,” *PLoS One* **14**(1), e0211060 (2019).
42. Q. N. Do et al., “MRI of the placenta accreta spectrum (PAS) disorder: radiomics analysis correlates with surgical and pathological outcome,” *J. Magn. Reson. Imaging* **51**(3), 936–946 (2020).
43. K. T. Leitch et al., “Placenta accreta spectrum and hysterectomy prediction using MRI radiomic features,” *Proc. SPIE* **12033**, 120331I (2022).
44. J. D. Dormer et al., “CascadeNet for hysterectomy prediction in pregnant women due to placenta accreta spectrum,” *Proc. SPIE* **12032**, 120320N (2022).
45. M. M. Fraz et al., “An ensemble classification-based approach applied to retinal blood vessel segmentation,” *IEEE Trans. Biomed. Eng.* **59**(9), 2538–2548 (2012).
46. R. Arnaout et al., “An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease,” *Nat. Med.* **27**(5), 882–891 (2021).
47. A. Kumar et al., “An ensemble of fine-tuned convolutional neural networks for medical image classification,” *IEEE J. Biomed. Health Inform.* **21**(1), 31–40 (2016).
48. J. Huang et al., “Topography-based feature extraction of the human placenta from prenatal MR images,” *Proc. SPIE* **12464**, 1246420 (2023).
49. P. Fabian, “Scikit-learn: machine learning in Python,” *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
50. G. Ke et al., “LightGBM: a highly efficient gradient boosting decision tree,” in *Adv. Neural Inf. Process. Syst.* **30**, pp. 3149–3157 (2017).
51. G. Klambauer et al., “Self-normalizing neural networks,” in *Adv. Neural Inf. Process. Syst.* **30**, pp. 972–981 (2017).
52. J. Motl and P. Kordík, “Stratified cross-validation on multiple columns,” in *IEEE 33rd Int. Conf. Tools Artif. Intell. (ICTAI)*, pp. 26–31 (2021).
53. D. Hendrycks and K. Gimpel, “Gaussian error linear units (GELUs),” arXiv:1606.08415 (2016).
54. J. He et al., “ReLU deep neural networks and linear finite elements,” arXiv:1807.03973 (2018).

Vishnu Vardhan Reddy Kanamata Reddy is a graduate research assistant in the Quantitative Bioimaging Laboratory (www.fei-lab.org) at the University of Texas at Dallas. He received his MS degree in computer science from the University of Texas at Dallas in 2023. His current research interests include deep learning, biomedical imaging, and computer vision.

Baowei Fei is a professor of bioengineering and Cecil H. and Ida Green Chair in Systems Biology Science at the University of Texas at Dallas. He is the director of the Quantitative Bioimaging Laboratory (www.fei-lab.org) and the director of the Center for Imaging and Surgical Innovation at University of Texas at Dallas and UT Southwestern Medical Center.

Biographies of the other authors are not available.