

# Uterine Ultrasound Image Captioning Using Deep Learning Techniques

Abdennour Boulesnane<sup>1\*,†</sup>, Boutheina Mokhtari<sup>2†</sup>, Oumnia Rana Segueni<sup>2†</sup> and Slimane Segueni<sup>3</sup>

<sup>1</sup>BIOSTIM Laboratory, Faculty of Medicine, Salah Boubnider University, Constantine, Algeria

<sup>2</sup>Department of IFA, Faculty of NTIC, Abdelhamid Mehri University, Constantine, Algeria

<sup>3</sup>Obstetrics and Gynecology Clinic, 23 Khelifi Abderrahmane Street, Chelghoum Laid, Mila, Algeria

## Abstract

Medical imaging has revolutionized medical diagnostics and treatment planning, progressing from early X-ray usage to sophisticated methods like MRIs, CT scans, and ultrasounds. This paper investigates the use of deep learning for medical image captioning, with a particular focus on uterine ultrasound images. These images are crucial in obstetrics and gynecology for diagnosing and monitoring various disorders across diverse age demographics. Nonetheless, their interpretation frequently proves difficult because of their intricacy. In this paper, a deep learning-based medical image interpretation system is developed, which integrates convolutional neural networks with bidirectional recurrent unit networks. This hybrid methodology examines both textual and visual components to produce relevant captions for ultrasound images of the uterus. The experimental findings demonstrate the efficacy of this strategy relative to baseline procedures, as indicated by superior BLEU and ROUGE scores. The suggested approach has superior performance in generating precise and informative captions. Our research enhances the interpretation of uterine ultrasound images, enabling physicians to make prompt and precise diagnoses, thereby elevating patient care.

## Keywords

Medical Image Captioning, Deep Learning, Uterine Ultrasound Images, Image Interpretation, Medical AI, Diagnostic Precision

## 1. Introduction

Throughout history, technological advances in medical imaging have dramatically changed the way we diagnose and treat [1]. This change began with the invention of X-rays over a century ago, which allowed imaging of the human body without the need for surgery. The field has since continued to evolve with new technologies such as magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET), and ultrasound, which have helped in accurately diagnosing many medical conditions [2]. Today, the integration of computer vision, natural language processing (NLP), and artificial intelligence (AI) has revolutionized the field, opening the door to unprecedented advances [3].

AI technologies, especially those based on deep learning, have shown remarkable potential in diagnosing various medical conditions quickly and accurately [4]. These machine learning algorithms significantly reduce the workload of medical professionals, leading to significant impacts on healthcare and patient care [5]. One of the most exciting developments in this field is medical image annotation (MIC) [6]. By leveraging deep learning, medical image annotation systems can automatically generate medical image annotations, combining expert annotations with images from comprehensive datasets to provide accurate and detailed analyses. These capabilities enhance medical documentation, speed up diagnosis, and facilitate remote consultations, thereby improving healthcare delivery overall [7].

---

TACC'2024, The 4th Tunisian-Algerian Conference on applied Computing, December 17-18, 2024, Constantine, Algeria

\*Corresponding author.

†These authors contributed equally.

✉ aboulesnane@univ-constantine3.dz (A. Boulesnane)

🌐 <https://aboulesnane.net> (A. Boulesnane)

🆔 0000-0002-2272-4953 (A. Boulesnane)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Despite these advances, the interpretation of medical images remains a formidable challenge [7]. Variations in physicians' levels of expertise can lead to inconsistent diagnoses, and misinterpretations can lead to medical errors that negatively impact patients' health. Furthermore, reading and analyzing these images can be time-consuming, especially in emergency settings where rapid decision-making is critical to the patient's life. Uterine ultrasound images, in particular, present unique challenges in obstetrics and gynecology. Their generally low quality compared to other medical images complicates the interpretation process, which can lead to delayed or incorrect diagnoses and impact patient care [8]. The complexity and diversity of these images underscore the need for an effective MIC system, which is the primary motivation for our research.

Our study aims to address the challenges in interpreting uterine ultrasound images by developing a specialized MIC system to enhance diagnostic accuracy and efficiency. To this end, we collected a comprehensive dataset of uterine ultrasound images, prioritizing patient privacy and confidentiality. This dataset was then carefully annotated using expert-provided descriptions, ensuring high-quality data for training and evaluation. We then performed extensive data preprocessing, isolating regions of interest within the images using a cropping algorithm and standardizing text captions using natural language processing techniques.

In the feature extraction phase, we used pre-trained convolutional neural network (CNN) models such as Inception V3 and DenseNet201 to obtain more detailed feature vectors from the images. Meanwhile, we converted the text data into numerical representations to match the image features. Our deep learning model combines these processed inputs through a bidirectional gated recurrent unit (BiGRU) network, generating descriptive captions for ultrasound images. Evaluated using metrics such as BLEU and ROUGE scores, the CNN-BiGRU model showed promising results in accurately describing uterine ultrasound images. These results demonstrate the effectiveness of our approach and its potential to enhance diagnostic accuracy in gynecology, ultimately contributing to improved patient care.

The remainder of the paper is structured as follows: Section 2 presents a review of related works. Section 3 details the proposed approaches. In Section 4, we analyze and discuss the experimental results. Finally, Section 5 offers conclusions and outlines directions for future research.

## 2. Related Work

Ultrasound imaging is invaluable for visualizing complex anatomical structures, offering advantages such as portability, real-time imaging, cost-effectiveness, and the absence of radiation [9]. However, interpreting these images can be challenging due to their often low quality, with common issues such as fuzzy borders and numerous artifacts [10]. While numerous studies have focused on medical image captioning (MIC) [6], the majority target medical reports for chest X-ray images [11], leaving MIC for ultrasound images relatively underexplored. This section will delve into MIC research specifically pertaining to ultrasound images.

In [10], a coarse-to-fine ensemble model for ultrasound image captioning is presented. The model first detects organs using a coarse classification model, then encodes the images with a fine-grained classification model, and finally generates annotation text describing disease information using a language generation model. The model, trained using transfer learning from a pre-trained VGG16 model, achieves high accuracy in ultrasound image recognition.

Building on the concept of combining different models, [12] introduces an NLP-based method to caption fetal ultrasound videos using vocabulary typical of sonographers. This approach combines a CNN (based on VGGNet16, fine-tuned on fetal ultrasound images) and an RNN for textual feature extraction. The CNN extracts image features, while the RNN encodes text features, merging them to generate captions for anatomical structures. The model is evaluated with BLEU and ROUGE-L metrics and produces relevant and descriptive captions for educating sonography trainees and patients.

In [13], a new method for ultrasound image captioning based on region detection is introduced to improve disease content analysis. The model detects and encodes focus areas in ultrasound images and then uses LSTM to generate descriptive text. This method increases accuracy in focus area detection and

achieves higher BLEU-1 and BLEU-2 scores with fewer parameters and faster runtimes than traditional models.

Expanding on incorporating additional data types, [14] introduces a Semantic Fusion Network to improve the accuracy of medical image diagnostic reports by integrating pathological information. This network comprises a lesion area detection model that extracts visual and pathological data and a diagnostic generation model that combines this information to produce reports. This method enhances the accuracy of generated reports, showing a 1.2% increase in the ultrasound image dataset compared to models relying solely on visual features.

In a similar vein of enhancing multimodal integration, [15] introduces an Adaptive Multimodal Attention network to generate high-quality medical image reports. The model employs a multilabel classification network to predict local properties of ultrasound images, using their word embeddings as semantic features. It integrates semantic and adaptive attention mechanisms with a sentinel gate to balance focus between visual features and language model memories. This approach enhances report accuracy and robustness, outperforming baseline models in capturing key local properties.

Addressing the challenge of small datasets, [16] presents a weakly-supervised method to enhance image captioning models using a large anatomically-labeled image classification dataset. This encoder-decoder model generates pseudo-captions for unlabeled images, creating an augmented dataset that significantly improves fetal ultrasound image captioning. This approach nearly doubles BLEU-1 and ROUGE-L scores, saving time on manual annotations and improving model performance in communicating information to laypersons.

In [17], a transformer-based model is proposed to generate descriptive ultrasound images of lymphoma, providing auxiliary guidance for sonographers. The model integrates deep stable learning to eliminate feature dependencies and includes a memory module for enhanced semantic modeling. Using a nonlinear feature decorrelation method, this approach visualizes cross-attention for interpretability and focuses on lymphoma features over the background. The result is a more accurate and detailed depiction of lymphoma in ultrasound images.

To further improve automatic report generation, [18] introduces a framework utilizing both unsupervised and supervised learning to align visual and textual features. Unsupervised learning extracts knowledge from text reports, guiding the model, while a global semantic comparison mechanism ensures accurate, comprehensive reports. Tested on three large datasets (breast, thyroid, liver), the method outperforms other approaches without needing manual disease labels, enhancing efficiency and accessibility.

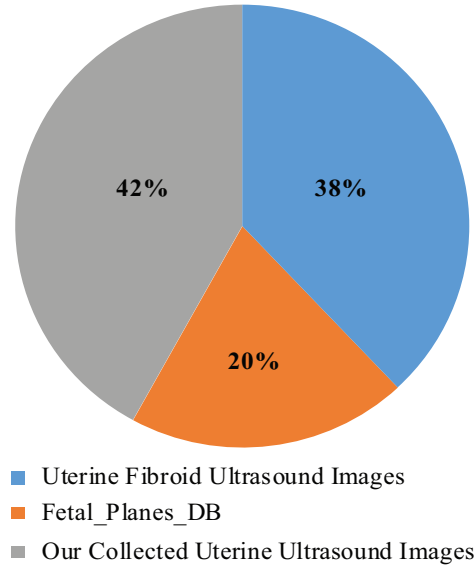
### **3. Methodology and Proposed Approach**

This study presents a novel uterine ultrasound image captioning system. To achieve this, we first gathered a diverse dataset of uterine ultrasound images and carefully annotated them with precise medical terminology, covering women of various ages and pregnancy stages. Our approach involved data preprocessing for images and text, followed by feature extraction using pre-trained CNN-based models. Finally, we implemented our proposed deep learning model, CNN-BiLGRU. Detailed descriptions of each module follow in the subsequent sections.

#### **3.1. Data Collection and Annotation**

Our dataset focuses specifically on gynecology, the branch of medicine that deals with women's health. We created a dataset that delves deeper into the details of gynecological imaging to address the specific challenges doctors face when diagnosing gynecological problems. This section details collecting and annotating medical images to train the medical image captioning model.

Our research utilized a dataset of ultrasound images exceeding 500 in number (505 images). Data collection involved acquiring ultrasound images from three main sources (see Figure 1). Internally, we gathered 214 images obtained directly from the Sonoscape SS1-8000 machine. Each image has a dimension of 1024x768 pixels (width: 1024 pixels, height: 768 pixels) and is stored in JPG format. Externally, we incorporated data from publicly available datasets to enrich this collection



**Figure 1:** Proportion of used uterine ultrasound images by source.

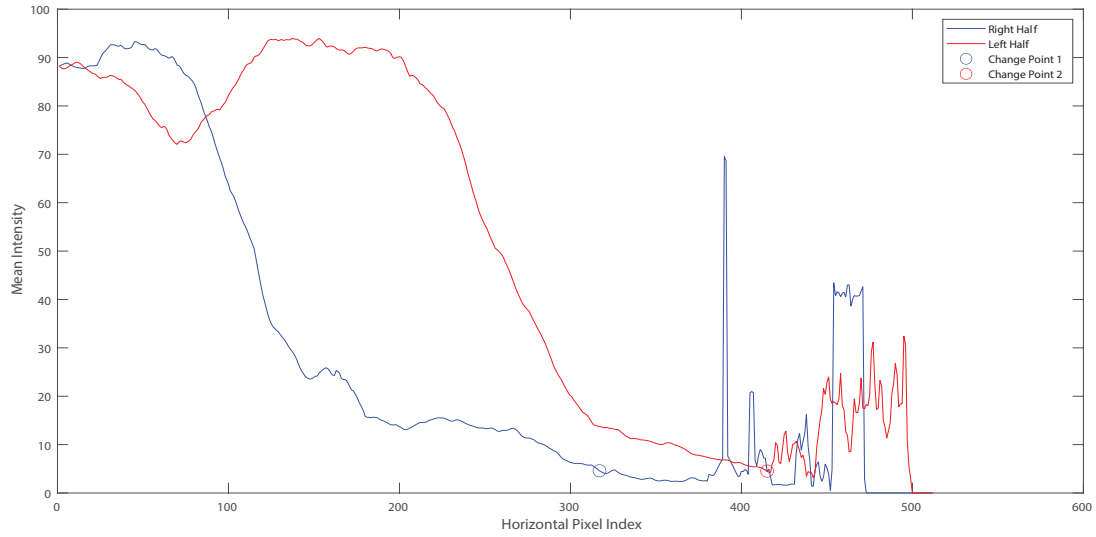
and capture a wider range of variations. From the Mendeley repository, which offered a rich collection of nearly 1,500 fetal ultrasound images (uterine fibroid ultrasound images [19]), we collaborated with experts to meticulously review and select a subset of 191 images that best aligned with our research goals. This selection process involved eliminating images with repetitive features, poor capture of the region of interest, or other factors that could negatively impact model training. Additionally, the Zenodo [20] dataset (Fetal Planes DB) provided 450 images, meticulously organized to include four images per patient, each representing the standard fetal planes of the abdomen, brain, femur, and thorax. From this collection, we selected a subset of 100 images that best aligned with our research goals, ensuring comprehensive coverage of fetal anatomy across multiple datasets.

In the form of captions, annotations were then added to each image in our dataset. These captions captured key features and findings within the ultrasound images, including identifying anatomical structures such as the stomach, umbilical vein, femur bones, and brain ventricles and noting potential abnormalities such as dilated organs or fluid pockets in the brain. We communicated closely with experts during the annotation process to ensure accuracy and quality. This collaboration helped us resolve image-related issues and made our dataset more valuable for analysis and research.

### 3.2. Data Pre-processing

Data preprocessing is crucial for ensuring the quality and usability of data for subsequent analysis and modeling [21]. Our study encompasses rigorous processing of images and text to enhance data integrity and relevance.

Image processing plays a crucial role in refining collected data. Initially, we analyze and filter the images to align with project requirements. The first step involves cropping the images to focus on the Region of Interest (ROI). Upon reading each ultrasound image, we convert it to grayscale if it is in color. Subsequently, we determine the cropping points by identifying significant changes in pixel intensity from the image center toward its edges. This process begins by calculating the mean intensity column-wise for both the right and left halves of the image, as depicted in Figure 2a. Peaks in these intensity profiles highlight areas of interest, and points where intensity drops below a predefined threshold (5% of peak value) denote edges of the ROI (see Figure 2b). Using these change points, we derive precise cropping coordinates to isolate the ROI (Figure 2c). Post-cropping, all images are resized uniformly to 224x224 pixels, a standard size compatible with many pre-trained neural networks. Each resized image instance is then converted into a Numpy array and normalized. Normalization involves scaling pixel values from 0 to 255 to a normalized range of 0 to 1 by dividing each pixel value by 255.



(a) Mean intensity variation across image width.



(b) Original image with crop lines.



(c) ROI-cropped image.

**Figure 2:** Original image to the region of interest.

After completing the image processing and preparing the images, we focused on processing the text captions associated with each image. These captions were derived from expert-provided medical descriptions and were systematically linked to their corresponding image file names within an Excel file. To enhance the text data for subsequent analysis, we applied NLP techniques [22]:

- **Convert to Lowercase:** All sentences were converted to lowercase to maintain consistency and reduce variability across the dataset.
- **Remove Punctuation:** Punctuation marks were systematically removed to simplify the text and emphasize the words.
- **Remove Single Letters:** Single letters such as 'l', 's', 'a', and 'à' were removed, as they typically do not contribute significant meaning in medical contexts.
- **Remove Extra Spaces:** Any extraneous spaces within the text were eliminated to ensure uniform spacing and improve text clarity.
- **Add Start and End Tags:** Special tags <START> and <END> were appended to the beginning and end of each sentence. These tags serve as markers during subsequent text processing and modeling to delineate sentence boundaries effectively.



### 3.3. Feature Extraction

Feature extraction is crucial in identifying and describing pertinent information within patterns [23]. This process facilitates pattern classification by establishing a structured and systematic approach. This phase focuses on deriving meaningful numerical representations from text descriptions and ultrasound images.

Text feature extraction involves converting textual data, such as medical reports and captions, into a format suitable for machine learning models. Initially, we employ a tokenizer to create a dictionary of word indices from our text data. This step allows us to determine the vocabulary size, represent the total number of unique words in the dataset, and identify the longest caption's length. Subsequently, we construct a vocabulary of unique words to map each word to its corresponding index. Shorter sequences are padded with zeros to ensure uniform input sequence lengths (captions), as neural networks require consistent input dimensions.

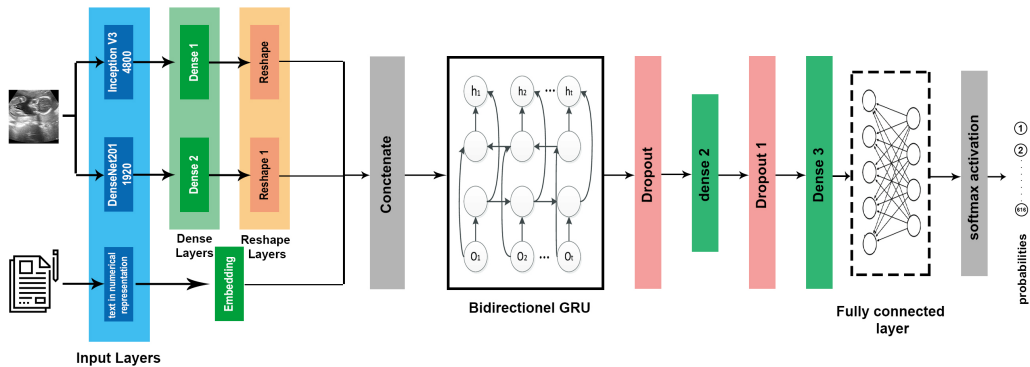
In addition to text, image feature extraction in this study utilizes advanced convolutional neural network architectures, namely Inception V3 and DenseNet201. These models have been pre-trained on the vast ImageNet dataset [24], which consists of millions of annotated images across thousands of categories. The key advantage of using these pre-trained models is their ability to capture intricate patterns and hierarchical representations within images.

### 3.4. Proposed Uterine Ultrasound Image Captioning Model

The proposed uterine ultrasound image captioning model aims to generate meaningful and accurate captions for medical ultrasound images of the uterus. To achieve high accuracy, the system architecture incorporates several advanced components. Our dataset consists of 505 images specifically selected to represent various uterine ultrasound scans commonly encountered in clinical practice.

As shown in Figure 3, the model's architecture begins with three input layers. The first input layer receives features extracted from a DenseNet201 model, shaped as (None, 1920). The second input layer obtains features from an InceptionV3 model, shaped as (None, 4800). The third input layer receives tokenized text sequences with a (None, 54) shape, where 54 represents the maximum caption length. The image features from the DenseNet201 and InceptionV3 models pass through dense layers that reduce their dimensionality to (None, 256). These outputs are then reshaped into (None, 1, 256) using reshape layers. Meanwhile, the tokenized text sequences are embedded, resulting in fixed-size tensor vectors (None, 54, 256).

The embeddings are concatenated with the reshaped image features, and the combined data is fed into a bidirectional GRU layer, which processes sequential data bidirectionally and produces an output shape of (None, 256). A dropout layer with a dropout rate of 0.5 is applied to prevent overfitting. The output is then passed through an intermediate dense layer that reduces the dimensionality to (None, 128), followed by another dropout layer with the same rate.



**Figure 3:** Architecture of the proposed CNN-BiGRU model.

Finally, a dense layer with a softmax activation function generates the final output, which has a shape of (None, 626). This represents the predicted caption probabilities for each word in the vocabulary. By integrating image and text features, this architecture produces accurate and informative captions for uterine ultrasound images, thereby enhancing medical diagnosis and treatment planning.

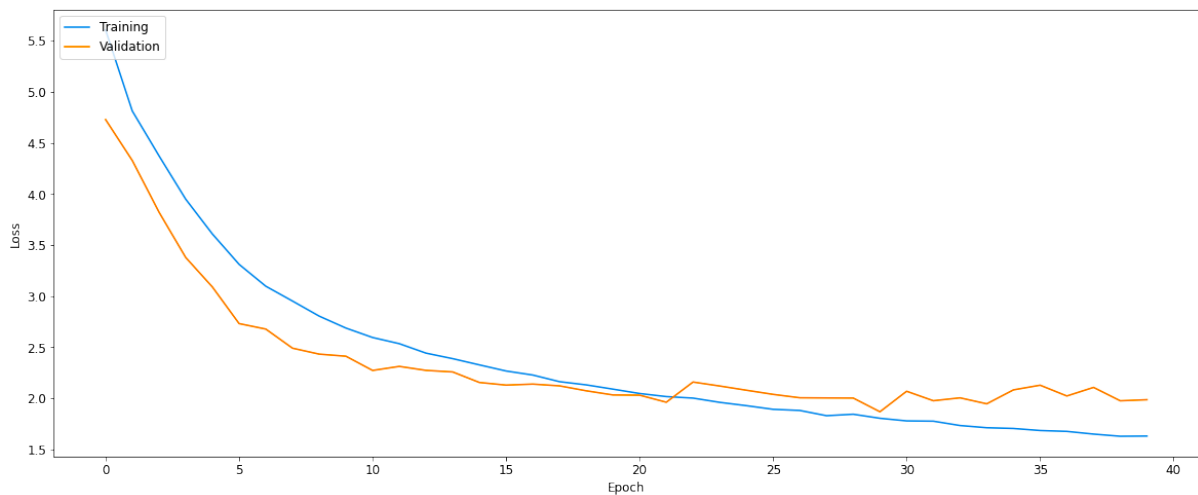
## 4. Experiments

In this section, we detail the experimental setup and analyze the results of our image captioning model. Regarding configuration, we divided the dataset, allocating 85% for training and 15% for testing (validation). Furthermore, the model parameters were configured with the Adam optimizer, a batch size of 16, and an early stopping patience of 10 epochs.

In this analysis, we evaluate the performance of the proposed model and compare it with other models to ensure the accuracy of the evaluation. To test the captions generated by the model, we used several metrics such as BLEU and ROUGE scores. BLEU scores (BLEU1, BLEU2, BLEU3, and BLEU4) are commonly used in machine translation to measure the similarity between generated labels and accurate references using n-grams. ROUGE scores (ROUGE1, ROUGE2, and ROUGEL) are used to evaluate text summarization, where ROUGE1 and ROUGE2 measure the recall of singletons and binaries. In contrast, ROUGEL evaluates the recall of the longest common subsequences between generated labels and references. These metrics help evaluate the model's ability to generate accurate and relevant captions for uterine ultrasound images.

### 4.1. Performance Analysis of the Proposed CNN-BiGRU Model

This analysis evaluates the performance of a CNN-BiGRU model that leverages powerful feature extraction capabilities using pre-trained Inception V3 and DenseNet201 architectures. Additionally, the model builds on BiGRU's strength in temporal sequence modeling, which helps it generate accurate and context-appropriate captions for uterine ultrasound images.



**Figure 4:** Loss curve of our proposed CNN-BiGRU model.

Figure 4 displays the learning curves for the loss of the proposed CNN-BiGRU model during both the training and validation phases. These curves indicate that the model was trained appropriately, with no signs of overfitting. The training and validation losses were closely aligned throughout the training process, a positive indicator of the model's generalization capability.

Specifically, at epoch 29, the model achieved a training loss of 1.64 and a validation loss of 1.86. These loss values suggest that the model effectively learned the underlying patterns in the data while maintaining a balance between fitting the training data and generalizing it to unseen validation data. The training

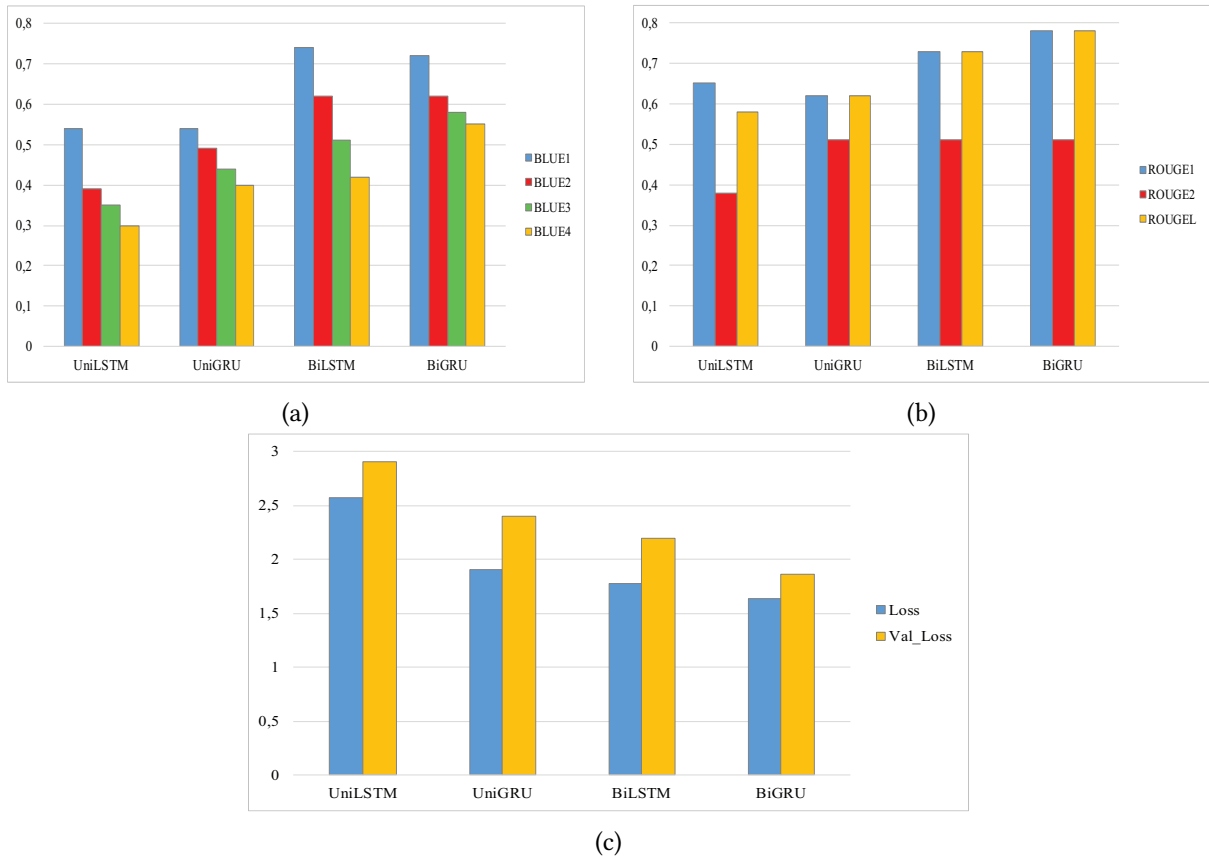
was terminated at epoch 39 due to the early stopping criterion, set with a patience of 10 epochs. This means that the model stopped training when there was no significant improvement in the validation loss for 10 consecutive epochs, thereby preventing overfitting and ensuring that the model maintained its performance on the validation set.

Achieving a low loss in image captioning is challenging because it requires understanding visual content, recognizing objects and relationships, and translating this into coherent text. Variability in descriptions and sequential dependency in caption generation add complexity. Additionally, aligning visual features with textual representations involves bridging the gap between two different data modalities (i.e., images and text).

## 4.2. Comparison with Baseline Models

Our study explored various architectures integrated with different processing layers to generate captions for uterine ultrasound images. We primarily focused on using DenseNet201 and InceptionV3 models for feature extraction, followed by BiGRU, as well as baseline models such as Unidirectional GRU (UniGRU), Bidirectional Long Short-Term Memory (BiLSTM), and Unidirectional Long Short-Term Memory (UniLSTM) networks.

To provide a comprehensive comparison, we evaluated the performance of these models using several metrics, including BLEU and ROUGE scores. Higher scores indicate better performance. As depicted in Figure 5a, the BLEU scores for our models showed that BiGRU and BiLSTM outperformed the baseline models UniGRU and UniLSTM, with BiGRU achieving the highest BLEU-4 score of 0.55. At the same time, the ROUGE scores highlighted BiGRU as the best performer, with a ROUGE-L score of 0.78, as


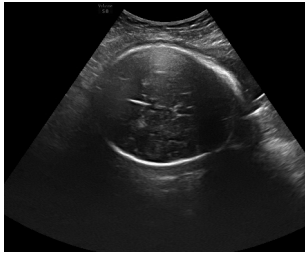


**Figure 5:** Comparative performance analysis of the proposed model and the baseline models. (a) Comparison of BLEU scores. (b) Comparison of ROUGE scores. (c) Comparison of training and validation loss.



**Table 1**

Sample outputs comparing reference captions with captions generated by the proposed model.

Uterine Ultrasound Image	Reference Caption	Generated Caption
	a white straight line at the top center that represents the femur bone it is possible to calculate the femur length the knee is straight	a white line at the top center that represents the femur bone it is possible to calculate the femur length the knee is straight
	a large slightly oval circle that represents the cranial contour of the fetus inside it is possible to see the cavum of the septum pellucidum on the right only but it is possible to calculate the biparietal diameter	a large slightly oval circle that represents the cranial contour of the fetus inside it is possible to see the cavum of the septum pellucidum can be seen on the right it is possible to calculate the biparietal diameter

shown in Figure 5b.

We also analyzed the training loss and validation loss for the selected models (see Figure 5c). The values of "Loss", which refers to the training loss calculated on the training dataset, and "Val\_Loss", which stands for validation loss calculated on the dataset, are important indicators of model performance. Our results showed that the BiGRU model achieved the lowest loss values (as shown in Figure 4), with a training loss of 1.64 and a validation loss of 1.86. This indicates the robustness and effectiveness of the model in generating accurate and context-appropriate annotations of uterine ultrasound images.

BiGRU's superiority lies in its ability to capture dependencies in both directions within sequences, a key feature for understanding context and generating accurate and consistent annotations. Unlike unidirectional models, BiGRU can process data in both forward and backward directions, providing a more comprehensive understanding of temporal context. This feature makes BiGRU particularly suitable for complex tasks such as generating image labels, where different parts of an image need to be accurately linked to their corresponding text while maintaining contextual consistency between them.

The CNN-BiGRU model outperformed the other models in terms of BLEU and ROUGE scores and showed lower loss values, proving its effectiveness in this application. In addition, Table 1 provides further evidence by comparing the reference comments with the comments generated by the model. This comparison clearly shows the model's ability to generate high-quality comments thanks to its bidirectional processing capabilities.

## 5. Conclusion and Future Work

In this study, we successfully developed a deep learning-based medical image interpretation system specifically designed for uterine ultrasound images using the CNN-BiGRU architecture. Our model effectively combined the image feature extraction capabilities of pre-trained CNNs (InceptionV3 and DenseNet201) with the sequential processing power of a bidirectional recurrent unit network. Through experimental study, this hybrid approach demonstrated superior performance over baseline models, achieving higher BLEU and ROUGE scores and maintaining low training and validation losses. The resulting captions were accurate and informative, improving the interpretability of complex uterine ultrasound images.

Our research findings demonstrate the potential of deep learning techniques to enhance diagnostic

accuracy and efficiency in obstetrics and gynecology. By automating the translation process, our model helps medical professionals make accurate and timely diagnoses and provide helpful second opinions, potentially improving patient outcomes.

The CNN-BiGRU model has shown good results, and there are ideas for further development in the future. One important one is to expand the dataset by adding ultrasound images of the uterus from different sources, which would make the model more robust and accurate. New techniques, such as attention mechanisms or transformer-based models, could also be tried to improve the quality of interpretations. In addition, work could be done to develop a system that annotates images in real-time for use in clinics. Creating user-friendly interfaces with feedback and integrating medical data from different sources could help provide more comprehensive diagnostic tools.

## Data Availability

The corresponding researchers could provide the data supporting the study's conclusions upon request.

## Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Grammarly to rephrase and perform Grammar and spelling checks. After using these tools, the authors reviewed and edited the content as needed. The authors take full responsibility for the publication's content.

## References

- [1] M. A. Haidekker, *Medical Imaging Technology*, Springer New York, 2013. doi:10.1007/978-1-4614-7073-1.
- [2] W. G. Bradley, History of medical imaging, *Proceedings of the American Philosophical Society* 152 (2008) 349–361.
- [3] R. Obuchowicz, M. Strzelecki, A. Piórkowski, *Artificial Intelligence in Medical Imaging and Image Processing*, MDPI, 2024. doi:10.3390/books978-3-7258-1260-8.
- [4] K. Suzuki, Overview of deep learning in medical imaging, *Radiological Physics and Technology* 10 (2017) 257–273. doi:10.1007/s12194-017-0406-5.
- [5] D. Kaul, H. Raju, B. K. Tripathy, *Deep Learning in Healthcare*, Springer International Publishing, 2021, p. 97–115. doi:10.1007/978-3-030-75855-4\_6.
- [6] D.-R. Beddiar, M. Oussalah, T. Seppänen, Automatic captioning for medical imaging (mic): a rapid review of literature, *Artificial Intelligence Review* 56 (2022) 4019–4076. doi:10.1007/s10462-022-10270-w.
- [7] L. Xu, Q. Tang, J. Lv, B. Zheng, X. Zeng, W. Li, Deep image captioning: A review of methods, trends and future challenges, *Neurocomputing* 546 (2023) 126287. doi:10.1016/j.neucom.2023.126287.
- [8] B. Levienaise-Obadia, A. Gee, Adaptive segmentation of ultrasound images, *Image and Vision Computing* 17 (1999) 583–588. doi:10.1016/s0262-8856(98)00177-2.
- [9] H. Chen, Y. Zheng, J.-H. Park, P.-A. Heng, S. K. Zhou, *Iterative Multi-domain Regularized Deep Learning for Anatomical Structure Detection and Segmentation from Ultrasound Images*, Springer International Publishing, 2016, p. 487–495. doi:10.1007/978-3-319-46723-8\_56.
- [10] X.-H. Zeng, B.-G. Liu, M. Zhou, Understanding and generating ultrasound image description, *Journal of Computer Science and Technology* 33 (2018) 1086–1100. doi:10.1007/s11390-018-1874-8.
- [11] X. Wang, G. Figueredo, R. Li, W. E. Zhang, W. Chen, X. Chen, A survey of deep learning-based radiology report generation using multimodal data, 2024. URL: <https://arxiv.org/abs/2405.12833>. doi:10.48550/ARXIV.2405.12833.

- [12] M. Alsharid, H. Sharma, L. Drukker, P. Chatelain, A. T. Papageorgiou, J. A. Noble, Captioning Ultrasound Images Automatically, Springer International Publishing, 2019, p. 338–346. doi:10.1007/978-3-030-32251-9\_37.
- [13] X. Zeng, L. Wen, B. Liu, X. Qi, Deep learning for ultrasound image caption generation based on object detection, *Neurocomputing* 392 (2020) 132–141. doi:10.1016/j.neucom.2018.11.114.
- [14] X. Zeng, L. Wen, Y. Xu, C. Ji, Generating diagnostic report for medical image by high-middle-level visual information incorporation on double deep learning models, *Computer Methods and Programs in Biomedicine* 197 (2020) 105700. doi:10.1016/j.cmpb.2020.105700.
- [15] S. Yang, J. Niu, J. Wu, Y. Wang, X. Liu, Q. Li, Automatic ultrasound image report generation with adaptive multimodal attention mechanism, *Neurocomputing* 427 (2021) 40–49. doi:10.1016/j.neucom.2020.09.084.
- [16] M. Alsharid, H. Sharma, L. Drukker, A. T. Papageorgiou, J. A. Noble, Weakly Supervised Captioning of Ultrasound Images, Springer International Publishing, 2022, p. 187–198. doi:10.1007/978-3-031-12053-4\_14.
- [17] J. Deng, D. Chen, C. Zhang, Y. Dong, Generating lymphoma ultrasound image description with transformer model, *Computers in Biology and Medicine* 174 (2024) 108409. doi:10.1016/j.compbimed.2024.108409.
- [18] J. Li, T. Su, B. Zhao, F. Lv, Q. Wang, N. Navab, Y. Hu, Z. Jiang, Ultrasound report generation with cross-modality feature alignment via unsupervised guidance, 2024. doi:10.48550/ARXIV.2406.00644.
- [19] T. Yang, Uterine fibroid ultrasound images, 2023. doi:10.17632/n2zcmcygb.2.
- [20] X. P. Burgos-Artizzu, D. Coronado-Gutierrez, B. Valenzuela-Alcaraz, E. Bonet-Carne, E. Eixarch, F. Crispi, E. Gratacós, FETAL\_PLANES\_DB: Common maternal-fetal ultrasound images, 2020. doi:10.5281/zenodo.3904280.
- [21] A. Boulesnane, S. Meshoul, K. Aouissi, Influenza-like illness detection from arabic facebook posts based on sentiment analysis and 1d convolutional neural network, *Mathematics* 10 (2022) 4089. doi:10.3390/math10214089.
- [22] A. Boulesnane, Y. Saidi, O. Kamel, M. M. Bouhamed, R. Mennour, Dzchatbot: A medical assistant chatbot in the algerian arabic dialect using seq2seq model, in: 2022 4th International Conference on Pattern Analysis and Intelligent Systems (PAIS), IEEE, 2022. doi:10.1109/pais56586.2022.9946867.
- [23] A. O. Salau, S. Jain, Feature extraction: A survey of the types, techniques, applications, in: 2019 International Conference on Signal Processing and Communication (ICSC), IEEE, 2019. doi:10.1109/icsc45622.2019.8938371.
- [24] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, *Communications of the ACM* 60 (2017) 84–90. doi:10.1145/3065386.
- [25] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
- [26] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.