Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 5 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i05.54331

PAPER

Image Captioning for Medical Surveillance in Smart Home Environments Using Vision Transformers

Lamiae Eloutouate[1](✉), Hicham Gibet Tani[2], Fatiha Elouaai[1], Mohammed Bouhorma[1], Mohamed Walid Hajoub[3]

[1]FSTT, Abdelmalek Essaadi University, Tetouan, Morocco

[2]FPL, Abdelmalek Essaadi University, Tetouan, Morocco

[3]ENSATE, Abdelmalek Essaadi University, Tetouan, Morocco

lamiae.eloutouate@uae.ac.ma

ABSTRACT

Medical surveillance in smart homes represents a transformative approach to patient care by utilizing advancements in computer vision to monitor and analyze patient behavior continuously. This study builds upon previous research by fine-tuning vision transformer (ViT) neural networks with a curated dataset that includes diverse scenarios of patients in both normal and abnormal conditions. The proposed model generates descriptive captions from surveillance camera images, effectively capturing contextual information and identifying potential medical indicators. These insights are integrated into an automated notification system designed to alert healthcare providers promptly, enabling timely and informed interventions. To evaluate the effectiveness of the approach, the fine-tuned ViT model is compared against traditional convolutional neural networks (CNNs) state-of-the-art model, demonstrating superior performance with an accuracy of 87.2%, a BLEU-4 score of 0.351, and a ROUGE-2 score of 0.591. These results highlight the model's ability to generate accurate and contextually relevant captions, outperforming CNN-LSTM baselines in accuracy, robustness, and contextual understanding. The findings underscore the critical role of artificial intelligence (AI) in detecting changes in patient conditions and providing personalized care through real-time monitoring. This proof-of-concept highlights the feasibility of deploying AI-driven solutions in medical surveillance systems, paving the way for innovative healthcare technologies. By addressing key challenges in patient monitoring, the study establishes ViT as a reliable and scalable tool for enhancing the quality and efficiency of healthcare delivery in smart home environments.

KEYWORDS

vision transformers (ViT), medical surveillance, image captioning, smart healthcare, artificial intelligence (AI) in healthcare

1 INTRODUCTION

The integration of artificial intelligence (AI) into healthcare has opened new frontiers for patient care, particularly in the field of medical surveillance. In recent years, the development of smart home technologies has enabled continuous monitoring

Eloutouate, L., Tani, H.G., Elouaai, F., Bouhorma, M., Hajoub, M.W. (2025). Image Captioning for Medical Surveillance in Smart Home Environments Using Vision Transformers. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(5), pp. 113–126. https://doi.org/10.3991/ijoe.v21i05.54331

Article submitted 2025-01-09. Revision uploaded 2025-02-13. Final acceptance 2025-02-13.

of patients in non-clinical environments, offering significant advantages for managing chronic conditions, elderly care, and post-operative recovery. By leveraging computer vision, healthcare providers can gain insights into patient behavior, detect deviations from normal routines, and take timely action to mitigate risks [1].

A promising approach to this challenge is the application of image captioning techniques to surveillance data, enabling automated analysis of patient behavior. In a previous study, we explored the use of vision transformer (ViT) neural networks pre-trained on the COCO dataset to generate descriptive captions from surveillance images, comparing their performance with traditional convolutional neural networks (CNNs) [2]. The findings demonstrated the superiority of ViT in generating natural language descriptions, offering a foundation for AI-driven monitoring systems. However, the reliance on generalized datasets like common objects in context (COCO) posed limitations in capturing healthcare-specific contexts.

Building on this foundation, the present study fine-tunes the ViT model using a curated dataset representing patients in both normal and abnormal situations. By generating descriptive captions and analyzing medical indicators, the model aims to alert healthcare providers to potential issues. This research not only underscores the role of AI in enhancing personalized patient care but also addresses key challenges in deploying robust and accurate medical surveillance systems.

Furthermore, we compare the performance of the fine-tuned ViT model with state-of-the-art computer vision models, evaluating metrics such as accuracy, contextual understanding, and robustness. By presenting a proof-of-concept model, this study seeks to establish the viability of ViT as a cornerstone for future healthcare innovations.

To address the challenges of medical surveillance in smart home environments, this study explores how fine-tuned ViT models can enhance the accuracy and contextual relevance of image captioning for patient monitoring. Specifically, we investigate how the proposed ViT-based model compares to traditional CNN-LSTM architectures in detecting and describing abnormal patient behaviors. Furthermore, we examine how generated captions can be leveraged to identify medical indicators and trigger real-time alerts for healthcare providers. Finally, this study aims to highlight key challenges in deploying AI-driven image captioning for medical surveillance and propose potential solutions for improving model robustness and real-world applicability.

2 LITERATURE REVIEW

2.1 Overview of AI in healthcare

The integration of AI into healthcare has revolutionized patient care, particularly in diagnostics, treatment planning, and continuous monitoring. AI-driven systems have demonstrated significant potential in analyzing complex medical data, enabling early detection of diseases, and providing personalized care [3]. In the context of medical surveillance, AI technologies such as computer vision and IoT sensors have been employed to monitor patient behavior in real-time, offering a proactive approach to managing chronic conditions and post-operative recovery [4]. Recent advancements in deep learning have further enhanced the accuracy and reliability of these systems, making them indispensable tools for modern healthcare [5].

However, while wearable devices and IoT sensors have been widely adopted for real-time monitoring [6], visual surveillance remains underexplored, particularly in translating raw video data into actionable clinical insights [7]. This gap highlights

the need for advanced AI models capable of generating contextually rich descriptions from visual data [8], which can complement existing sensor-based systems for comprehensive patient monitoring.

2.2 Image captioning in medical contexts

Vision transformers have emerged as a powerful alternative to traditional CNNs in image analysis tasks. By leveraging self-attention mechanisms, ViTs excel at capturing long-range dependencies in visual data, making them particularly suited for complex tasks such as medical image analysis [9].

Recent studies have explored ViTs' effectiveness across various medical applications. A comprehensive review by authors in [10] highlights the use of ViTs in medical imaging tasks, including classification, segmentation, detection, and clinical report generation. The study emphasizes ViTs' ability to model long-range dependencies, which is essential for analyzing complex anatomical structures in medical images.

In disease detection, ViTs have demonstrated superior performance. A study published in [11] applied ViTs to corneal confocal microscopy images for the automated detection of diabetic peripheral neuropathy. The model effectively captured both local and global features, achieving an AUC of 0.99, surpassing previous CNN-based methods. Similarly, authors in [12] developed a vision-series transformer (ViST) for screening coronary heart diseases using coronary CT angiography, achieving an accuracy of 83.78%, outperforming CNN-based methods.

In the same context, integrating ViTs with large language models has shown promising results. Authors of the paper [13] proposed a method guided by the segment anything model (SAM), which enhances encoding through a combination of general and detailed feature extraction. Their approach outperformed the pre-trained BLIP2 model in generating descriptive captions for medical images, showcasing the potential of ViTs in generating highly accurate medical descriptions.

In another paper [14], authors demonstrated ViTs' superiority in detecting early-stage tumors in MRI scans. However, their potential for medical image captioning remains largely untapped, with only a few studies [15] exploring ViTs for generating diagnostic reports. Despite their promise, the application of ViTs in medical surveillance systems remains limited, highlighting a critical gap in the literature. This gap presents an opportunity to leverage ViTs for generating descriptive captions in medical surveillance, particularly in smart home environments where real-time monitoring is critical.

2.3 Gaps in current research

While significant progress has been made in applying AI to healthcare, several challenges remain. First, the reliance on generalized datasets limits the ability of models to capture healthcare-specific contexts. For instance, models trained on public datasets struggle to generalize to medical scenarios. Second, existing systems often lack the robustness required to handle diverse patient scenarios, particularly in non-clinical environments. For example, CNN-LSTM architectures, while effective in general captioning tasks, often fail to capture nuanced patient behaviors in medical settings. Third, there is a need for fine-tuned models that can generate accurate and actionable insights from medical surveillance data. Addressing these gaps is essential for developing reliable and scalable AI-driven healthcare solutions.

The proposed study aims to bridge these gaps by fine-tuning a ViT model on a curated dataset of cardiac patient behaviors, enabling the generation of contextually rich captions for real-time medical surveillance.

3 MATERIALS AND METHODS

3.1 Research design

This study aims to fine-tune a ViT model for generating descriptive captions from surveillance images in smart home environments, specifically for monitoring cardiac patients as a case study. The approach leverages advancements in computer vision and natural language processing to identify medical indicators and integrate them into an automated notification system for healthcare providers. The research design addresses the limitations of generalized datasets and improves the accuracy of medical surveillance systems.

3.2 Dataset description

The dataset used in this study relies on a model that was pretrained on the COCO dataset [16], which is a large-scale dataset for image recognition, including tasks such as object detection, segmentation, and annotation. It contains over 330,000 images annotated with 80 object categories and five captions per image. It is widely used in computer vision research.

Training a model from scratch for this task can be highly resource-intensive and computationally expensive. Additionally, relying on a general public dataset may result in a model that lacks the specificity needed for medical applications, as such datasets often fail to capture critical domain-specific nuances. Therefore, fine-tuning a pretrained model on a specialized dataset is preferred, as it not only leverages existing knowledge but also enhances the model's ability to learn contextually relevant features, improving accuracy and performance for the targeted task (see Figure 1).

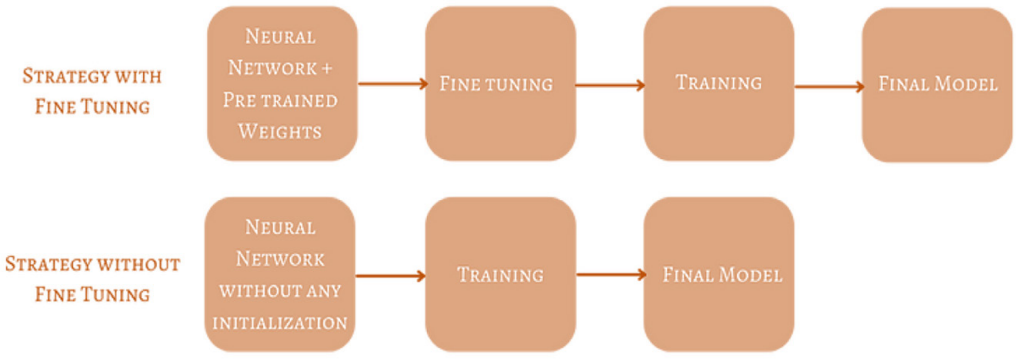

Fig. 1. Fine-tuning technique steps to create a new model

A customized dataset was gathered to monitor the behaviors of cardiac patients at home, using a case study approach. The dataset was carefully designed to cover a wide range of patient scenarios, ensuring diversity in patient demographics, environments, and lighting conditions to improve model generalization.

The dataset consists of 1100 images categorized into two main types of behaviors:

1. **Normal behaviors:** Activities that do not pose an immediate health risk and do not require intervention. These include:
   a) Taking prescribed medications
   b) Engaging in light physical activity (e.g., walking, exercising)
   c) Watching TV or resting
2. **Abnormal behaviors:** Activities that indicate potential health risks and require urgent intervention. These include:
   a) Vertigo and fainting
   b) Stress and Fatigue (excessive tiredness, lack of energy)
   c) Chest pain (clutching chest, signs of distress)
   d) Stomach pain
   e) Smoking (holding or using tobacco products)
   f) Falls (tripping, falling, lying motionless)

Each image in the dataset was manually annotated ensuring high-quality labels and medical relevance. The annotation process involved:

- Assigning five descriptive captions per image, capturing key contextual details related to the observed behavior.
- Validation by multiple annotators to ensure consistency and reduce bias.
- Focus on abnormal behaviors to enhance the model's ability to detect critical medical conditions in high-risk patients.

Additionally, efforts were made to balance the dataset by including varied patient postures, different home environments, and multiple camera angles to improve real-world applicability. Figure 2 presents a sample of this dataset, illustrating the diversity of the collected images.

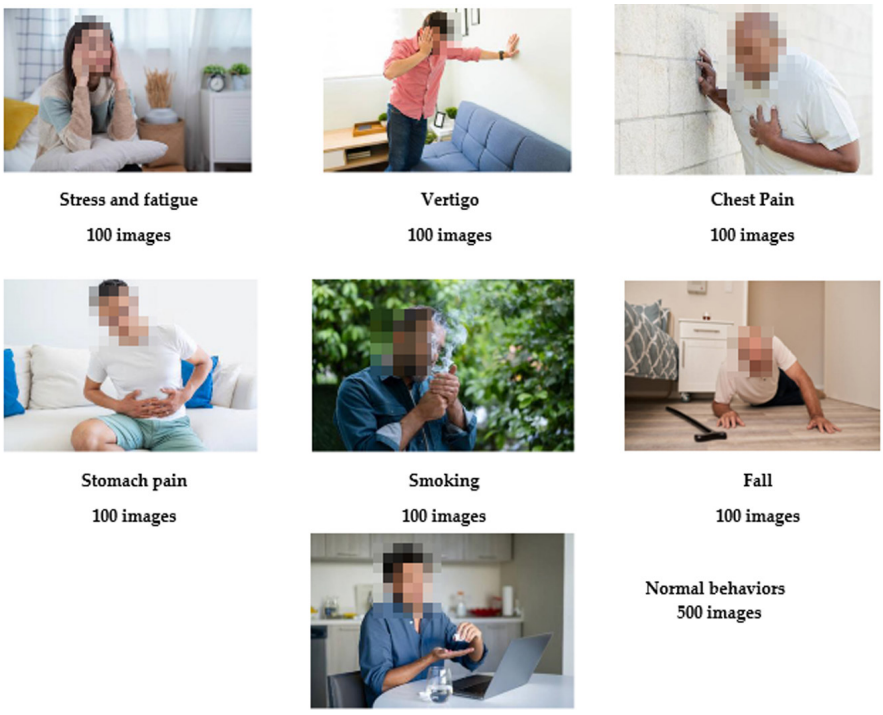

Fig. 2. Sample dataset used for fine-tuning the proposed model

Data augmentation techniques were applied to improve model robustness, including:

- Random cropping and resizing to 224 × 224 pixels.
- Horizontal flipping to simulate different viewpoints.
- Color jittering to account for variations in lighting conditions.
- Gaussian noise to enhance the model's ability to handle low-quality images.

3.3 Transformers architecture for image captioning

The ViT model was employed for image captioning, leveraging its encoder-decoder architecture. The process involves the following steps (see Figure 3):

1. Dividing the image into patches: The input image is resized to a fixed resolution $X \in RH \times W \times 3$ and divided into N patches, where $N = (H/P) \times (W/P)$ and P is the patch size (P = 16 in our setup).
2. Flattening and vectorizing patches: Each patch is flattened into a 1D sequence and vectorized into embeddings.
3. Adding positional encoding: A learnable 1D positional encoding is added to each patch feature to form the final encoder input Pa = [P1, ..., PN]
4. Encoding features with attention: The encoder applies self-attention mechanisms to capture relationships between patches, producing a feature matrix that highlights crucial spatial relationships.
5. Decoding features: The decoder generates captions based on the encoder's output. During training, it builds a vocabulary, and during prediction, it generates captions for new images.

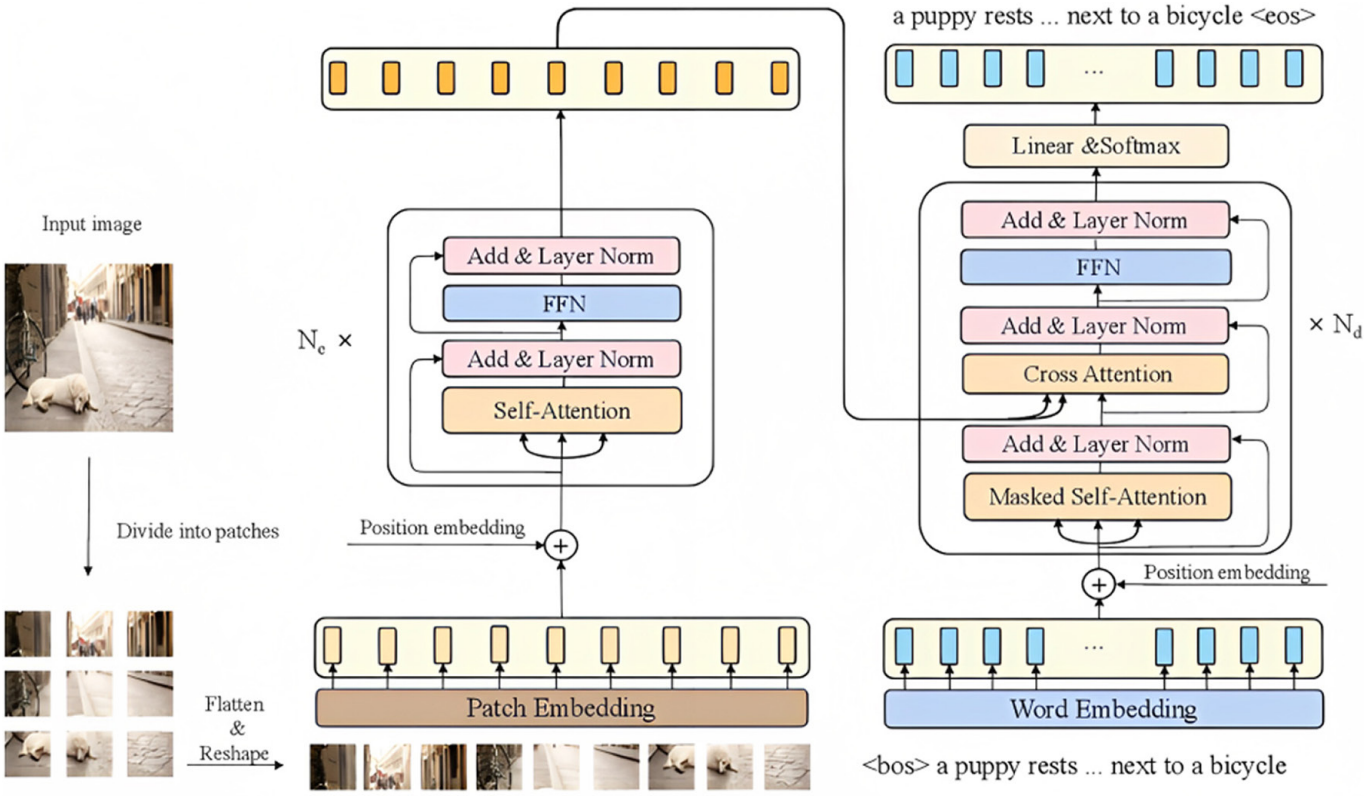

Fig. 3. Transformers architecture for generating image captions [17]

3.4 Model architecture

The architecture combines a "VisionEncoderDecoderModel," using ViT to encode the input image and GPT-2 to decode the encoded image into textual captions (see Figure 4).

First, the input image is preprocessed by resizing it and dividing it into fixed-size patches, which are then flattened and embedded into a higher-dimensional space. Positional encodings are added to retain spatial information. The ViT encoder processes these patches using multi-head self-attention mechanisms to extract contextual features. These features are fed into the GPT-2 decoder, which generates captions autoregressively by predicting one token at a time, conditioned on the image features and previously generated tokens. The model is fine-tuned on a custom dataset of cardiac patient images, freezing the lower layers of the pre-trained models and updating the upper layers to adapt to the specific task. This architecture enables the generation of accurate and contextually relevant captions for medical surveillance.

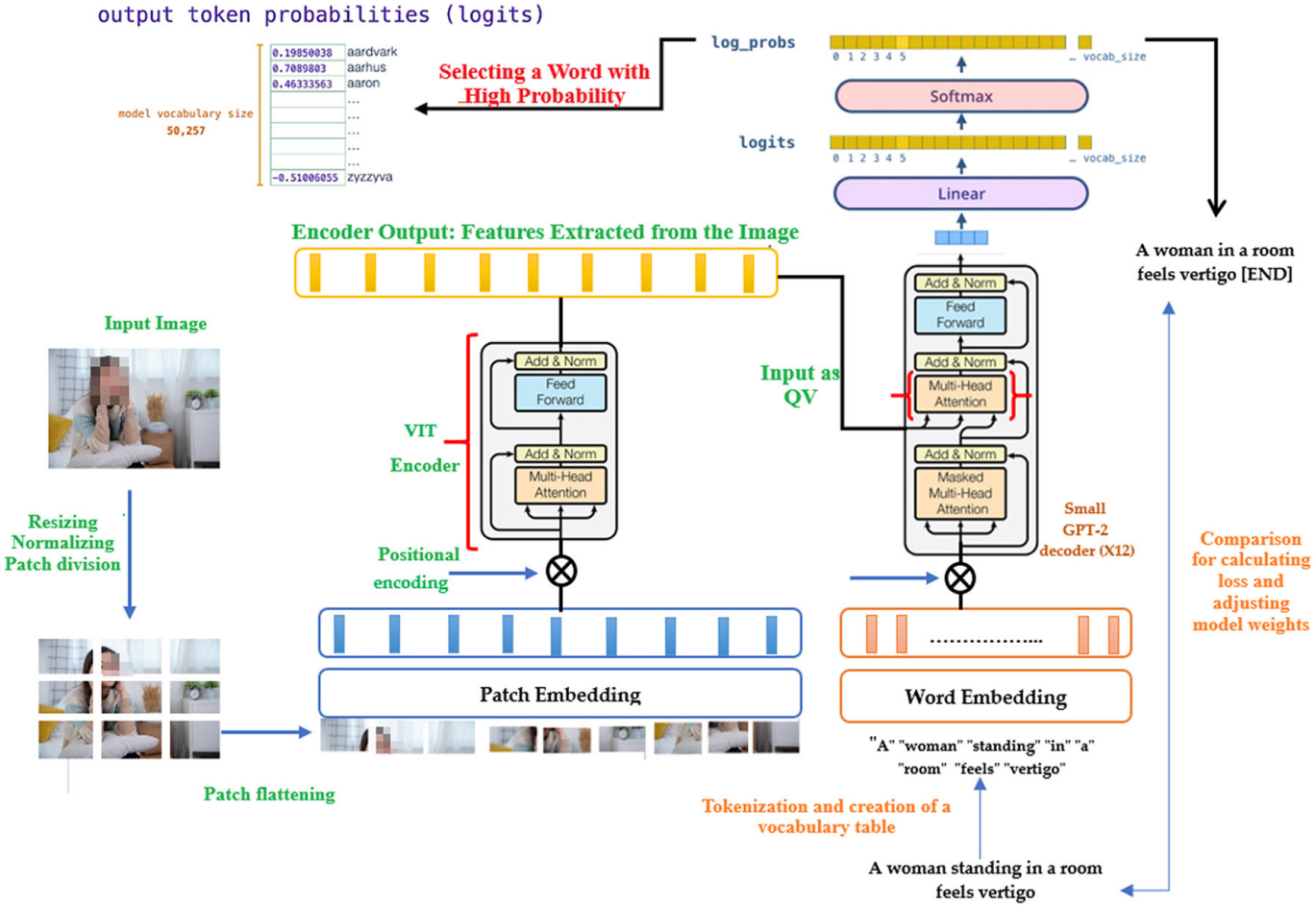

Fig. 4. The proposed ViT model architecture for caption generation

3.5 Training and evaluation

The model was trained using cross-entropy loss and the AdamW optimizer, with a learning rate of 1e-4 and a batch size of 32 for 50 epochs. Input images were resized to 224 × 224 pixels and normalized before training. A warm-up schedule was

used for the first 5% of training steps to stabilize learning. Additionally, the lower layers of the model were frozen to retain general feature extraction capabilities, while the upper layers were fine-tuned to adapt to the medical surveillance domain.

The GPT-2 decoder was fine-tuned on the annotated medical captions from the custom cardiac patient dataset to adapt it to the healthcare domain. During training, the model was initialized with weights pretrained on general text masses. A causal language modeling objective was employed, and teacher forcing was used to condition the model on ground truth captions. This domain adaptation process ensured that GPT-2 could generate coherent and contextually relevant medical captions, enhancing its suitability for healthcare applications.

To benchmark the proposed model's performance, traditional CNNs with LSTM were trained and evaluated on the same dataset, highlighting the effectiveness of the ViT architecture. CNN-LSTM was selected as a baseline due to its widespread use and established performance in image captioning tasks, particularly in healthcare applications. While advanced models like CLIP, Swin Transformer, and ViT-Hybrid models exist, CNN-LSTM provides a well-understood benchmark for evaluating the effectiveness of the proposed ViT + GPT-2 architecture. Additionally, CNN-LSTM's architecture, which combines convolutional layers for feature extraction with recurrent layers for sequence modeling, offers a clear contrast to the transformer-based approach, allowing us to highlight the advantages of self-attention mechanisms and fine-tuning strategies in medical captioning.

Model performance was evaluated using the following metrics:

- ROUGE-2: Measures the similarity between generated captions and reference captions, focusing on bigram matches to assess contextual alignment.
- BLEU: Evaluates text similarity by comparing n-grams between generated and reference captions, commonly used in machine translation tasks.

This evaluation framework initially ensures a comprehensive assessment of the model's ability to generate accurate and contextually relevant captions.

Future efforts will aim to optimize the model further by incorporating techniques such as learning rate decay, dropout, and regularization [18]. Specifically, we plan to apply learning rate decay to gradually reduce the learning rate during training, which will help achieve a more efficient convergence and prevent large updates in the later stages of training. To mitigate overfitting, we intend to explore the use of dropout and L2 regularization. Dropout will be applied to randomly deactivate a fraction of neurons during training, reducing the risk of co-adaptation and improving the model's generalization capabilities. Additionally, L2 regularization will be employed to penalize large weights, further helping to prevent overfitting. Furthermore, we plan to monitor validation loss trends more closely in future work to ensure effective learning. We expect that incorporating these strategies will help improve the model's ability to generalize and optimize its performance on unseen data.

4 RESULT AND DISCUSSIONS

4.1 Overview of experimental setup

The proposed system combines a ViT for image encoding and a GPT-2 model for caption generation, fine-tuned on a custom dataset of cardiac patient images (see Figure 2). The dataset includes both normal behaviors (e.g., taking medication,

exercising) and abnormal behaviors (e.g., breathlessness, chest pain). The model was trained using cross-entropy loss and the AdamW optimizer, with images resized to 224 × 224 pixels. For comparison, traditional CNN-LSTM models were also evaluated. Performance was assessed using ROUGE-2, BLEU, precision, and recall metrics.

4.2 Performance of the proposed model

The proposed ViT + GPT-2 model demonstrated superior performance compared to baseline models. Key results include:

Table 1. The proposed ViT model performance

Model	Error	Accuracy Score	BLEU	ROUGE-2
Transformers	0.128	0.872	0.351	0.591
CNN-LSTM	0.280	0.72	0.24	0.52

The proposed model achieved an accuracy score of 87.2%, significantly outperforming the CNN-LSTM baseline, which scored 0.72. This improvement indicates better alignment with reference captions and enhanced contextual understanding.

In terms of ROUGE-2 score, which evaluates the quality of generated text by measuring the overlap of bigrams (pairs of consecutive words) between the generated and reference text, the proposed model achieved nearly 60% precision, compared to the CNN-LSTM model's performance. Similarly, for the BLEU score, the proposed model achieved 35.1%, surpassing the CNN-LSTM baseline of 24%. This highlights the model's ability to generate more accurate and fluent captions.

These results underscore the effectiveness of the ViT + GPT-2 architecture in generating high-quality captions and identifying critical medical indicators.

In the same context, the proposed model generated contextually rich and accurate captions for both normal and abnormal patient behaviors. For instance, in an image depicting a patient experiencing chest pain (see Figure 5), the model generated the caption: "Person feeling pain in the heart." In contrast, the CNN-LSTM model produced a more generic caption: "A person is sitting on a couch."

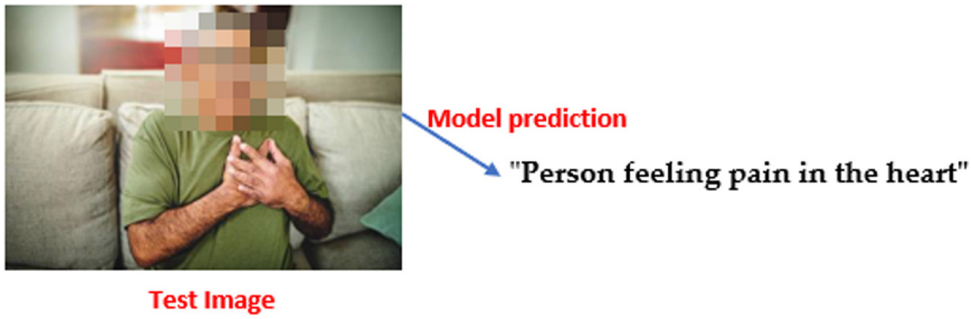

Fig. 5. The proposed ViT model prediction example

These examples highlight the proposed model's ability to capture nuanced details and provide actionable insights for healthcare providers. Moreover, to assess the statistical significance of the performance differences between the proposed ViT + GPT-2 model and the baseline CNN-LSTM model, we conducted independent t-tests on the BLEU and ROUGE-2 scores. The results showed that the improvements achieved

by the proposed model were statistically significant ($p < 0.01$ for both metrics). Additionally, we computed 95% confidence intervals for the accuracy scores, which were [85.5%, 88.9%] for the proposed model and [70.1%, 73.9%] for the CNN-LSTM baseline. These results confirm that the proposed model's superior performance is not due to random variation but reflects a meaningful improvement in medical captioning tasks.

4.3 Effectiveness of behavioral change detection

The system analyzes generated captions to detect behavioral changes through a series of steps. First, preprocessing is applied, where captions are converted to lowercase, tokenized, and stripped of stop words. Lemmatization is then performed to normalize the tokens. In the similarity evaluation step, the tokens are compared against a predefined list of medical indicators, such as breathlessness, fatigue, and chest pain. If a token match any of these medical indicators, an alert is generated and sent to healthcare providers through a dedicated communication tool.

This pipeline ensures accurate and timely detection of behavioral changes, enabling proactive interventions.

4.4 Real-time alert system

The system integrates the caption analysis pipeline with a real-time notification system to improve responsiveness. For example, when the system detects signs of breathlessness in a patient, it sends an alert to the healthcare provider, enabling immediate intervention. Alerts are delivered through multiple channels, such as a mobile app, email, or SMS, ensuring timely communication. This feature significantly enhances the system's practicality and impact in real-world healthcare settings.

4.5 Comparison with baseline models

The proposed ViT + GPT-2 model outperformed traditional CNN-LSTM models and other state-of-the-art approaches across all evaluation metrics. Key advantages include:

Table 2. LSTM and transformers comparison based on BLEU metric [19]

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Transformers	0.36	0.34	0.32	0.31
CNN+LSTM	0.26	0.27	0.18	0.17

- **Contextual understanding:** The ViT + GPT-2 architecture captures complex spatial relationships and generates more contextually relevant captions. For example, while the proposed model achieved a BLEU-2 score of 0.351 (Table 1), prior studies using CNN-LSTM architectures reported BLEU scores in the range of 0.17–0.27 (Table 2).
- **Robustness:** The model performs consistently across diverse scenarios, including low-light conditions and occluded images. In contrast, CNN-LSTM models often struggle with such variations, as noted in [20].

- **Scalability:** The system can be extended to other medical domains by fine-tuning on domain-specific datasets, whereas CNN-LSTM models require extensive retraining and architectural modifications [21].

These comparisons highlight the superiority of the proposed model and its potential to advance the field of medical image captioning.

4.6 Integration with IoT sensor data

The proposed system integrates IoT sensor data (see Figure 6) with visual analysis to enhance robustness and minimize the risk of false negatives, which is crucial for medical applications. For instance, in one case, abnormal heart rate data from IoT sensors triggered further investigation, even though no visible signs of distress were detected in the video feed. This multimodal approach provides comprehensive monitoring of patient health, addressing situations where visual cues alone may not reveal critical issues. By combining visual and sensor data, the system reduces the likelihood of missing important health events, making it more reliable and fault-tolerant for real-world medical use.

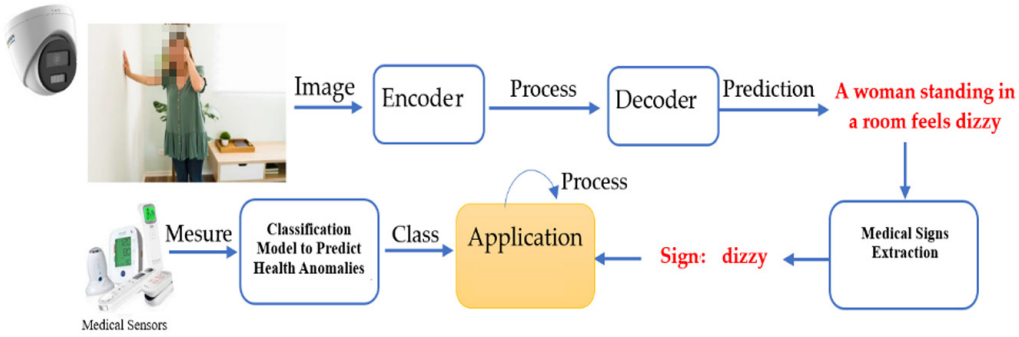

Fig. 6. Medical signs extraction and validation using medical sensors

4.7 Challenges and limitations

While the proposed system demonstrates promising results, several challenges remain. One major concern is computational resources, as training the GPT-2 decoder requires significant processing power, necessitating the use of high-performance GPUs. Additionally, the model's performance is highly dependent on the quality and diversity of the training dataset. Expanding the dataset to include a broader range of patient scenarios could enhance its generalizability. Ethical considerations also play a crucial role, particularly in ensuring patient privacy and data security when deploying the system in real-world settings.

Future research directions aim to address these challenges and further improve the system. One key avenue is the integration of multimodal approaches, combining visual, sensor, and textual data to enhance accuracy and robustness. Another important focus is interpretability, where developing tools to explain the model's predictions can increase transparency and trust among healthcare providers. Expanding the dataset remains a priority, incorporating a wider range of patient scenarios and medical conditions to improve the model's applicability.

It is important to note that this study is exploratory in nature and serves as a proof of concept. The images used are publicly available and were anonymized through pixilation to protect privacy. However, these images do not represent clinical data and are used solely for illustrative purposes. Further research involving controlled, ethically sourced datasets is recommended to validate and expand upon these findings.

5 CONCLUSION

This study presents a novel AI-driven system for real-time medical surveillance in smart home environments, leveraging a fine-tuned ViT and GPT-2 model for image captioning and behavioral change detection. The proposed architecture demonstrates superior performance over traditional CNN-LSTM models, achieving a ROUGE-2 score of 0.591 and a BLEU score of 0.351, alongside high accuracy score in identifying critical medical indicators. By generating contextually rich and accurate captions, the system enables healthcare providers to monitor patients effectively and intervene promptly when abnormalities are detected.

The integration of IoT sensor data further enhances the system's robustness, addressing scenarios where visual cues alone may fail to reveal critical health issues as highlighted by our past research works on this field [22]. This multimodal approach minimizes false negatives, ensuring comprehensive and reliable patient monitoring. The system's ability to deliver real-time alerts via a dedicated communication tool underscores its practicality and potential to transform healthcare delivery.

Despite its promising results, challenges such as computational resource requirements and dataset limitations remain. Future work will focus on expanding the dataset to include more diverse patient scenarios, exploring multimodal approaches that combine visual, sensor, and textual data, and improving the model's interpretability for clinical adoption. These advancements will further solidify the system's role as a scalable and reliable tool for enhancing patient care in smart home environments.

6 REFERENCES

[1] A. Thacharodi *et al.*, "Revolutionizing healthcare and medicine: The impact of modern technologies for a healthier future-A comprehensive review," *Health Care Sci.*, vol. 3, no. 5, pp. 329–349, 2024. https://doi.org/10.1002/hcs2.115

[2] H. G. Tani, L. Eloutouate, F. Elouaai, M. Bouhorma, and M. W. Hajoub, "Transforming healthcare: Leveraging vision-based neural networks for smart home patient monitoring," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 10, pp. 20–32, 2023. https://doi.org/10.3991/ijoe.v19i10.40381

[3] J. Bajwa, U. Munir, A. Nori, and B. Williams, "Artificial intelligence in healthcare: Transforming the practice of medicine," *Future Healthc. J.,* vol. 8, no. 2, pp. e188–e194, 2021. https://doi.org/10.7861/fhj.2021-0095

[4] F. Kitsios, M. Kamariotou, A. I. Syngelakis, and M. A. Talias, "Recent advances of artificial intelligence in healthcare: A systematic literature review," *Appl. Sci.*, vol. 13, no. 13, p. 7479, 2023. https://doi.org/10.3390/app13137479

[5] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017. https://doi.org/10.1016/j.media.2017.07.005

[6] I. Alihamidi, A. Deroussi, A. Addaim, and A. Ait Madi, "Revolutionizing healthcare: Convergence of IoT and open-source ERP systems in health information management," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 20, no. 9, pp. 83–98, 2024. https://doi.org/10.3991/ijoe.v20i09.48805

[7] M. A. Hassan, A. S. Malik, N. Saad, B. Karasfi, D. Fofi, and W. Sohail "Towards health monitoring in visual surveillance," in *2016 6th International Conference on Intelligent and Advanced Systems (ICIAS)*, Kuala Lumpur, Malaysia, 2016, pp. 1–6. https://doi.org/10.1109/ICIAS.2016.7824046

[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 3156–3164. https://doi.org/10.1109/CVPR.2015.7298935

[9] R. N. Gabriel, A. D. Elvira, and S. Emilio, "From vision to text: A comprehensive review of natural image captioning in medical diagnosis and radiology report generation," *Medical Image Analysis,* vol. 97, p. 103264, 2024. https://doi.org/10.1016/j.media.2024.103264

[10] R. Azad *et al.*, "Advances in medical image analysis with vision transformers: A comprehensive review," *arXiv preprint arXiv:2301.03505*, 2023.

[11] C. Ben Rabah, I. N. Petropoulos, R. A. Malik, and A. Serag, "Vision transformers for automated detection of diabetic peripheral neuropathy in corneal confocal microscopy images," *Front. Imaging*, vol. 4, 2025. https://doi.org/10.3389/fimag.2025.1542128

[12] K. Wang, H. Meng, and X. Wang, "Application of vision-series transformer in screening for coronary heart diseases using coronary CT angiography," in *Proceedings of the 2023 4th International Conference on Computing, Networks and Internet of Things (CNIOT '23)*, 2023, pp. 421–425. https://doi.org/10.1145/3603781.3603858

[13] Z. Zhang *et al.*, "Sam-guided enhanced fine-grained encoding with mixed semantic learning for medical image captioning," *arXiv preprint arXiv:2311.01004*, 2023.

[14] S. Takahashi *et al.*, "Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review," *J. Med. Syst.*, vol. 48, 2024. https://doi.org/10.1007/s10916-024-02105-8

[15] P. Arshi, A. K. Muhammad, Z. Rukhsana, A. Huma, A. Muhammad, and M. F. Muhammad, "Vision transformers in medical computer vision—A contemplative retrospection," *Engineering Applications of Artificial Intelligence,* vol. 122, p. 106126, 2023. https://doi.org/10.1016/j.engappai.2023.106126

[16] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context." https://cocodataset.org/

[17] L. Wei, C. Sihan, G. Longteng, Z. Xinxin, and L. Jing, "CPTR: Full transformer network for image captioning," *arXiv preprint arXiv:2101.10804*, 2021.

[18] D. A. Cadillo-Laurentt and E. A. Paiva-Peredo, "Histopathological image classification using convolutional neural networks for detection of metastatic breast cancer in lymph nodes," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 20, no. 2, pp. 31–45, 2024. https://doi.org/10.3991/ijoe.v20i02.46789

[19] C. Zouitni, M. A. Sabri, and A. Aarab, "A comparison between LSTM and Transformers for image captioning," in *Digital Technologies and Applications, ICDTA 2023,* in Lecture Notes in Networks and Systems, S. Motahhir and B. Bossoufi, Eds., vol. 669, 2023, pp. 492–500. https://doi.org/10.1007/978-3-031-29860-8_50

[20] H. Javed, S. El-Sappagh, and T. Abuhmed, "Robustness in deep learning models for medical diagnostics: Security and adversarial challenges towards robust AI applications," *Artif. Intell. Rev.*, vol. 58, 2025. https://doi.org/10.1007/s10462-024-11005-9

[21] J. Maurício, I. Domingues, and J. Bernardino, "Comparing vision transformers and convolutional neural networks for image classification: A literature review," *Appl. Sci.*, vol. 13, no. 9, p. 5521, 2023. https://doi.org/10.3390/app13095521

[22] L. Eloutouate, F. Elouaai, H. G. Tani, and M. Bouhorma, "Home automation and machine learning models for health monitoring," in *Proceedings of the 5th International Conference on Big Data and Internet of Things, BDIoT 2021*, in Lecture Notes in Networks and Systems, M. Lazaar, C. Duvallet, A. Touhafi, and M. Al Achhab, Eds., vol. 489, 2021, pp. 362–372. https://doi.org/10.1007/978-3-031-07969-6_27

7 AUTHORS

Lamiae Eloutouate holds a PhD in Computer Science and Artificial Intelligence and is a member of the Data & Intelligent Systems Team at the Faculty of Sciences and Technology of Tangier (FSTT), Abdelmalek Essaadi University, Tetouan, Morocco. Her research focuses on smart homes, remote healthcare, and smart healthcare technologies (E-mail: lamiae.eloutouate@uae.ac.ma).

Hicham Gibet Tani is an Associate Professor and a qualified research supervisor in the Computer Science Department at the Polydisciplinary Faculty of Larache (FPL), Abdelmalek Essaadi University, Tetouan, Morocco. He is also a member of the Data & Intelligent Systems Team. His research interests include cloud computing, big data, machine learning, and smart cities (E-mail: h.gibettani@uae.ac.ma).

Fatiha Elouaai is a full Professor at the Faculty of Sciences and Technology of Tangier (FSTT), Abdelmalek Essaadi University, Tetouan, Morocco. Her research spans bioinformatics, human-computer interaction, computer communications, and cybersecurity. She actively contributes to academic research, publications, and scientific events, driving innovation across multiple disciplines (E-mail: elouaaif@gmail.com).

Mohammed Bouhorma is a full Professor at the Faculty of Sciences and Technology of Tangier (FSTT), Abdelmalek Essaadi University, Tetouan, Morocco, with over 25 years of teaching and research experience. His expertise includes information security, security protocols, artificial intelligence, big data, and digital forensics. His research interests encompass cybersecurity, IoT, big data analytics, AI, smart city technologies, and serious games (E-mail: mbouhorma@uae.ac.ma).

Mohamed Walid Hajoub is a PhD candidate at the National School of Applied Sciences of Tetouan (ENSATE), Abdelmalek Essaadi University, Tetouan, Morocco. His research focuses on data science, machine learning, and computer engineering. He is dedicated to advancing knowledge in these fields through rigorous academic research (E-mail: mohamedwalidhajoub1@gmail.com).

8 DISCLAIMER

This study is a proof of concept and utilizes publicly available images that have been pixelated to remove all identifiable personal features. These images are used solely for illustrative purposes, sourced ethically, and do not contain any clinical or sensitive data.