Explainable AI for Endometriosis Diagnosis: A Dense Neural Network Approach with SHAP Interpretation

Fatade Oluwayemisi Boye*, Afolashade Oluwakemi KUYORO, Ernest Enyinnaya Onuiri

Department of Computer Science. Babcock University

DOI: https://doi.org/10.51244/IJRSI.2025.12030070

Received: 07 March 2025; Accepted: 12 March 2025; Published: 15 April 2025

INTRODUCTION

According to [2], Artificial Intelligence's revolutionary potential is bringing about a fundamental revolution in every aspect of the world, impacting everything from product development to medical diagnosis. It also has several subfields, particularly in the practice of medicine, including computer vision (CV), deep learning (DL), and machine learning (ML).

Women's quality of life is greatly impacted by endometriosis, a chronic gynecological disorder that can lead to infertility and discomfort [10]. Even though it is common, non-invasive diagnosis is still difficult. Researchers have used information from imaging, blood tests, genetics, and symptoms to investigate different Machine Learning (ML) methods. Methods such as LASSO regression and logistic regression have demonstrated potential [11]. Significant restrictions still exist, nonetheless, which impair patient outcomes and clinical adoption.

First, current research frequently relies on a small number of data sources or a single data type, which may cause important information that could improve diagnostic accuracy to be missed. For instance, research that only looks at symptom data may overlook important information from other sources. Second, many machine learning models used in contemporary research are "black boxes," which means that their decision-making procedures are opaque [12]. This lack of interpretability limits trust and prevents broad clinical adoption by making it challenging for medical professionals to comprehend the process used to generate diagnosis.

When using XAI to diagnose endometriosis, decision confidence and trustworthiness are greatly increased (Antoniadi et al., 2021). By having a thorough understanding of how an AI system makes a diagnosis, clinicians can make well-informed decisions regarding its application, which will increase the degree of trust and adaptability of the technology in the medical field. Based on the features it prioritizes, it can detect any bias in the model's predictions.

Several machine learning (ML) approaches such as logistic regression, LASSO regression, and U-Net models to mention a few have been explored for non-invasive diagnosis, leveraging data from symptoms, genetic markers, blood tests, and imaging techniques.

This study aims to address the limitations of symptoms as a single data source and lack of interpretability by using 3 structured data types and leverage on SHAP for model’s prediction interpretability.

METHODS

Data Set Overview

This dataset includes features such as patient demographics (age, gender, etc.), clinical history (previous diagnoses, family medical history, etc.), and symptom profiles (pain levels, specific symptom onset, and duration). The structured data was initially collected and stored in a structured format (CSV) containing

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue III March 2025

various patient attributes. Each row in the dataset represents an individual patient, with columns capturing specific clinical and demographic characteristics, in total there are 1208.

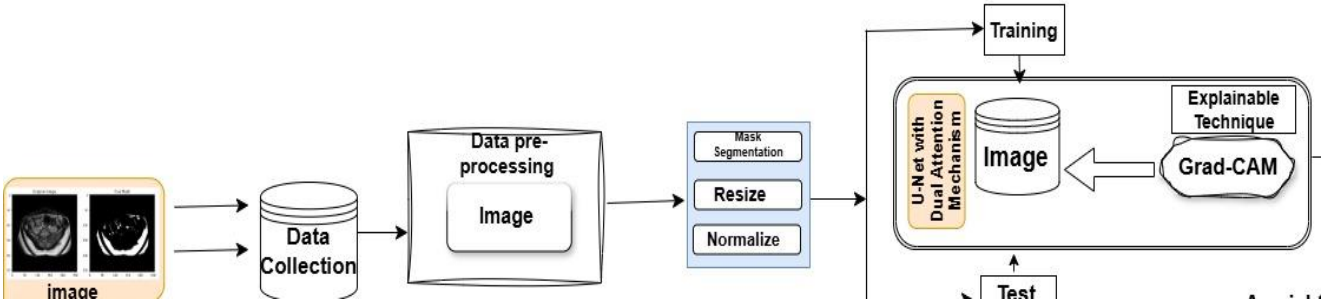

Figure 1. Framework for DNN integrated with SHAP. (Source: Researcher)

Data Preparation and Preprocessing

Data Cleaning was carried out by identifying missing values in critical fields (such as symptom severity or age) and handled by either filling them with median values (for numerical features) or using "Unknown" (for categorical data). A few outliers were noted and reviewed and adjusted or removed as necessary. Categorical variables, such as patient gender or categorical symptom levels, were encoded to numerical values using one-hot encoding techniques and label encoding (for ordinal data). This conversion was necessary to ensure compatibility with machine learning models that generally require numerical input. Numerical features, including age and symptom severity, were standardized to have a mean of zero and a standard deviation of one. This normalization was applied to maintain consistent data ranges and prevent any single feature from disproportionately impacting the model due to its scale. Because of the potential dimensionality of the dataset, a feature selection step was applied to retain only the most informative features. And correlation matrix visualization was used to visualize the relationships between one-hot encoded features in the non-image dataset, with values ranging from -1 (strong negative correlation) to 1 (strong positive correlation).

Data Organization

Data was organized into folders based on patient ID, symptoms, patient physical examination result, and pateint History. The dataset was split into 80% training, and 20% testing sets.

Model Selection (Dense Neural Network) and training

The data was used to train a dense neural network. The model iteratively learned from the data using backpropagation and optimization techniques. The training process was designed to optimize classification performance while maintaining interpretability. A dense feedforward neural network was implemented, consisting of multiple fully connected layers. ReLU was used for hidden layers, while a sigmoid activation function was applied in the output layer for binary classification. The Adam optimizer was used with a learning rate of 0.001 to ensure stable convergence. Binary cross-entropy was selected to measure classification performance. Training was conducted using a batch size of 32 over 50 epochs, with early stopping implemented to prevent overfitting.

This model was selected because of it’s ability to capture complex relationships between pateint attributes while maintaining flexibility and scalability. More importantly, it is compatible witrh SHAP, it allows detailed feature attribution. These ensures that model’s decision making process remains transparent and interpretable.

SHAP Integration

The Shapely Additive explanations (SHAP) method is based on Shapley values, a concept from cooperative game theory that ensures fair feature attribution. SHAP works with local explanations and global

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue III March 2025

explanations and provides both. It also gives consistent and theoretical explanations. It is based on game theory ensuring that feature attributions are mathematically consistent and fair. SHAP handles feature interaction, features like clinical symptoms, patient history may interact in complex ways, SHAP captures these features without interaction that may affect the overall result.

The mathematical expression for SHAP values is given by

$$\varphi_i = \sum_{scN\{i\}} \frac{|S|!\left(N|-|S|-1\right)!}{|N|!} |f(s \cup \{i\}) - f(s)|$$

Were

φ_i: SHAP value for feature I (its contribution to the prediction)

N: the set of all features.

$scN|\{i\}$: A subset of features excluding features i

$f(s)$: The model output when using only the features in S

$f(s \cup \{i\})$: The model output when adding features, I to subset S

The fraction weights each subset's contribution, ensuring fair allocation across all possible feature subset.

Tools

Python Programming Language

Python served as the primary programming language due to its versatility and extensive libraries for machine learning, medical imaging, and data analysis. With key libraries that include NumPy and Pandas for data manipulation and preprocessing of data, such as symptoms, clinical history, and demographic data, Matplotlib and Seaborn for visualizing data distributions and model evaluation metrics, Scikit-Learn for data preprocessing, splitting datasets, and evaluation metrics calculation.

Google Colab

Google Colab was used as the main development environment, providing GPU support for model training without the need for local high-performance hardware. It provides free GPU and TPU resources, ease of collaboration, integration with Google Drive, and a notebook interface for iterative development and documentation.

TensorFlow and Keras

TensorFlow and its high-level API, Keras, were utilized for building, training, and evaluating the U-Net model with a dual attention mechanism.

RESULTS AND DISCUSSION

Designing a Dense Model with SHAP

Data Preparation and Preprocessing

For consistency of dataset and to ensure quality, the required pre-processing steps were taken as recorded easlier to ensure a reliable model. Figure 3.1 and Figure 3.2 shows the visualization of the datasets after data cleaning, one-hot encoding, normalization, and feature selection.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue III March 2025

Figure 3.1. Visualization of dataset viewed per attribute

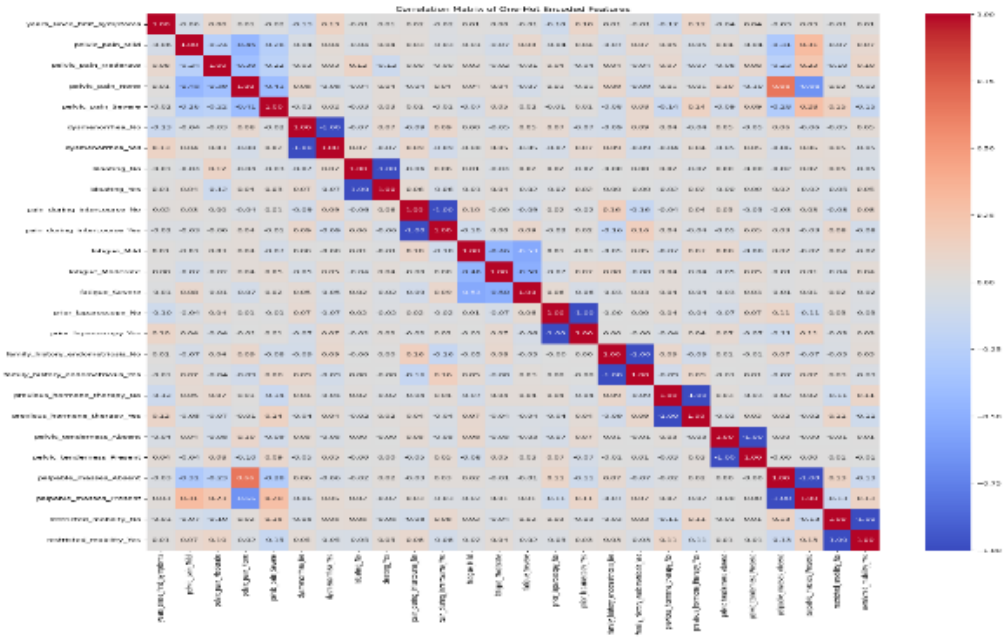

Figure 3.2. Visualization of correlation matrix of one-hot encoded features

The correlation matrix visualizes the relationships between one-hot encoded features in the non-image dataset, with values ranging from -1 (strong negative correlation) to 1 (strong positive correlation). From Figure 4.2, it is observed that there is minimal redundancy because most features exhibit low to moderate correlations with values that is closer to 0, it can be induced from this that no features need to be removed because of high correlation. The correlation result also shows that features like severality in pelvic pain and presence of palpable masses show moderate correlation about 0.66. this figure suggests that this relationship

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue III March 2025

might be clinically significant and it might be necessary to further investigate for potential interaction. There is absence of correlation close to 1 or -1 this indicates that there is no multicollinearity which can adversely affect the model performance. This result shows that we can retain all the features. The datasets were further scaled for easy analysis.

Data Splitting

The dataset was divided into training and testing sets to ensure robust model evaluation. 80% training, and 20% testing was applied to this as like MRI dataset splitting. Table 4.2shows the result of accurate splitting to training and testing sets.

Table 3.1. Results of non-image Training and Testing Set Splitting

Training set (80%)	966
Test set (20%)	242

Dence layer training

The non-image data was used to train the dense layer at the result as shown in Table 4.8a and Table 4.8b respectively. The classification report gives precision, recall, F1-score, and support for binary classification task. The two classes are labelled as “False” and “True”, these represents the negative and the positive cases, respectively. When the model predicts “False” which is a total of 59 samples class, it is correct 81% of the time, out of all the actual false cases, the model correctly identifies 87% of them. A balance between precision and recall was noted with F1-Score of 84% which indicates a strong classification performance. The model classified 183 samples as positive class, when it predicts “true” it is correct 87% of the time. Out of all the true cases, the model correctly identifies 81% of them as seen with the recall of 81% and F1-score of 83% shoes a strong balance of precision and recall for detecting positive cases.

The model correctly classifies 87% of all test samples, with Macro average: precision as 84% as average precision across both samples, recall at 84%, and F1-score as 83.5% across both classes. Weighted average precision of 86%, recall 87%, and F1-score at 87% balance across precision and Recall in each classes. It is generally observed that the model performance can be said to be good and reliable with accuracy of 87%

Integration with SHAP

Figure 4.8 shows SHAP explanation on Non-image data.

Longer menstrual cycles tend to increase the likelihood of endometriosis, while shorter cycles (blue) decrease it, Patients reporting left-side pelvic pain are more likely to be classified as having endometriosis, Older patients (red) tend to be at higher risk, while younger patients (blue) have a lower risk. Pelvic Pain Severity & Pain Intensity also indicates increased likelihood of diagnosis. The moderate Impact Features includes Pelvic Pain Frequency which is that Frequent pelvic pain is associated with a higher prediction for endometriosis, having undergone surgeries contributes to the model’s prediction, a known family history of endometriosis increases the probability of diagnosis, and urinary urgency & constipation. And Less Impactful Features included Dysmenorrhea Severity (Severe & Moderate) while still relevant, they have a lower impact compared to other pain-related features, Nodules in Cul de Sac which indicates the presence of nodules contributes to diagnosis but with lower overall impact, and Diarrhea (Yes).

Table 3.2a: Classification Report

	precision	Recall	F1-score	Support
FALSE	0.81	0.87	0.84	59
TRUE	0.87	0.81	0.83	183

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue III March 2025

Table 3.2b

Accuracy			0.87	206
macro Aveg	0.84	0.84	0.835	206
Weighted Avg	0.86	0.87	0.87	206

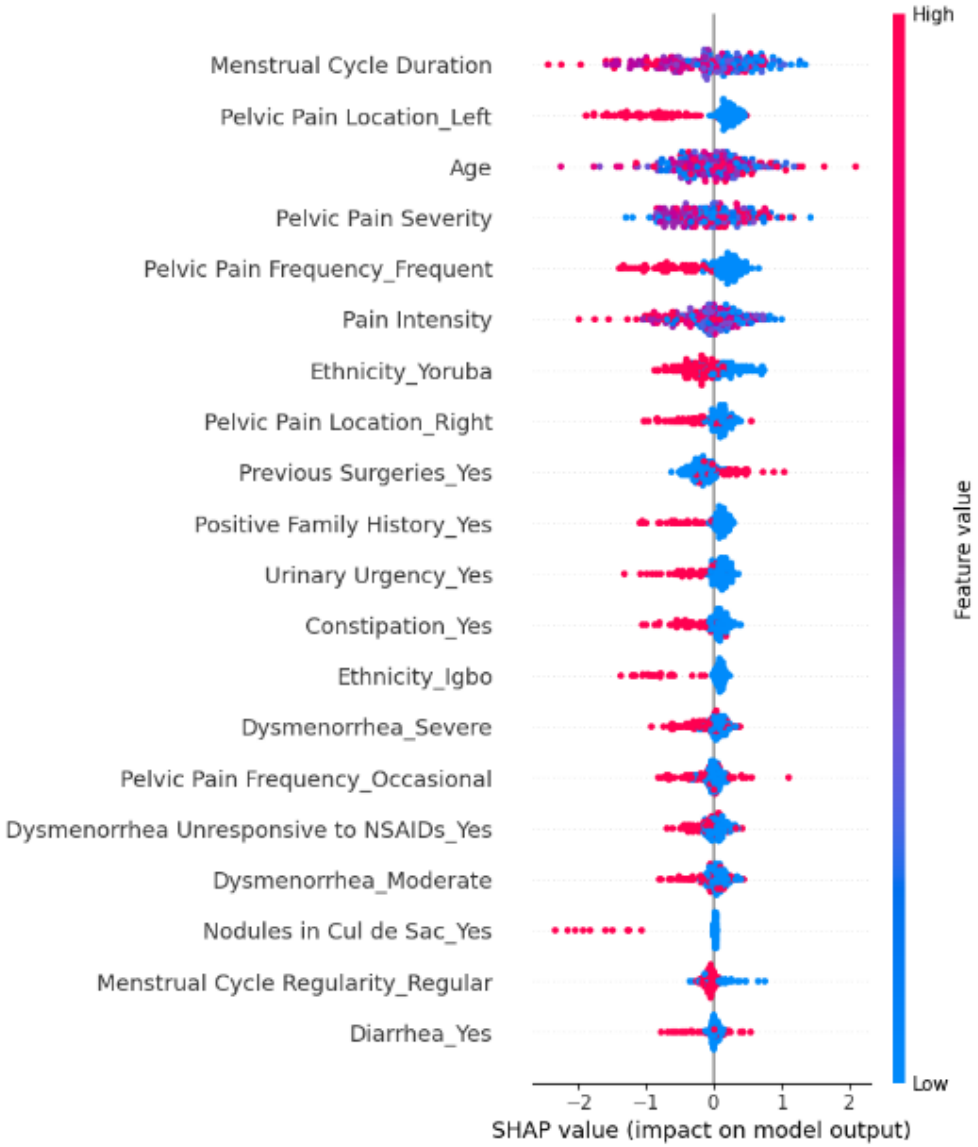

Figure 3.4. SHAP Explainability

DISCUSSION

The dataset was analyzed using a Dense Neural Network (DNN), leveraging structured patient data such as demographics, clinical history, symptoms profiles, and the result of physical examination for endometriosis diagnosis. Preprocessing steps ensured data quality through missing data handling, feature encoding, standardization, and feature selection.

The model's performance, evaluated using precision, recall, F1-Score, and accuracy, demonstrated strong classification ability. The overall accuracy of 87% indicates that the DNN effectively distinguished between positive and negative cases. The precision for positive cases (87%) and negative cases (81%) suggest that the model reliably predicts endometriosis while minimizing false positives. The recall scores 87% for negative and 0.81 for positive cases reflect the model's sensitivity to identify both conditions. The macro-average F1-score of 87% further confirms the model's balanced performance across classes.

To enhance interpretability, SHAP analysis was applied, identifying key features such as pain severity, previous diagnosis, and symptoms onset as primary contributors to model predictions. This feature attribution analysis provided transparency, allowing for clinical validation of the model's decision-making process.

CONCLUSIONS

DNN-based analysis of structured data produced reliable and interpretable results, demonstrating its potential as an independent diagnosis tool for endometriosis.

ACKNOWLEDGMENT

This work is supported by the many individuals who have made it a success.

REFERENCES

1. Wang et al., "Artificial intelligence in reproductive medicine," 2019, BioScientifica Ltd. doi: 10.1530/REP-18-0523.
2. E. Lutomski, S. Meaney, R. A. Greene, A. C. Ryan, and D. Devane, "Expert systems for fetal assessment in labour," Apr. 30, 2015, John Wiley and Sons Ltd. doi: 10.1002/14651858.CD010708.pub2.
3. Wang and A. Preininger, "AI in Health: State of the Art, Challenges, and Future Directions," Aug. 01, 2019, NLM (Medline). doi: 10.1055/s-0039-1677908.
4. Sivajohan, M. Elgendi, C. Menon, C. Allaire, P. Yong, and M. A. Bedaiwy, "Clinical use of artificial intelligence in endometriosis: a scoping review," Dec. 01, 2022, Nature Research. doi: 10.1038/s41746-022-00638-1.
5. Elgendi, C. Allaire, C. Williams, M. A. Bedaiwy, and P. J. Yong, "Machine Learning Revealed New Correlates of Chronic Pelvic Pain in Women," Front Digit Health, vol. 2, Dec. 2020, doi: 10.3389/fdgth.2020.600604.
6. Yoldemir, "Artificial intelligence and women's health," Jan. 02, 2020, Taylor and Francis Ltd. doi: 10.1080/13697137.2019.1682804.
7. Mu, M. Cui, and X. Huang, "Multimodal data fusion in learning analytics: A systematic review," Dec. 01, 2020, MDPI AG. doi: 10.3390/s20236856.
8. A. Lupean et al., "Differentiation of endometriomas from ovarian hemorrhagic cysts at magnetic resonance: The role of texture analysis," Medicina (Lithuania), vol. 56, no. 10, pp. 1–13, Oct. 2020, doi: 10.3390/medicina56100487.
9. Mao, C. Chen, H. Gao, L. Xiong, and Y. Lin, "A deep learning-based automatic staging method for early endometrial cancer on MRI images," Front Physiol, vol. 13, Aug. 2022, doi: 10.3389/fphys.2022.974245.
10. Bhardwaj et al., "Machine Learning for Endometrial Cancer Prediction and Prognostication," Jul. 27, 2022, Frontiers Media S.A. doi: 10.3389/ fonc.2022. 852746.
11. Yang et al., "Global, regional and national burden of anxiety disorders from 1990 to 2019: results from the Global Burden of Disease Study 2019," Epidemiol Psychiatr Sci, vol. 30, 2021, doi: 10.1017/S2045796021000275.
12. Kuznetsov, K. Dworzynski, M. Davies, and C. Overton, "Diagnosis and management of endometriosis: Summary of NICE guidance," BMJ (Online), vol. 358, 2017, doi: 10.1136/bmj.j3935.
13. Kaul, S. Enslin, and S. A. Gross, "History of artificial intelligence in medicine," Oct. 01, 2020, Mosby Inc. doi: 10.1016/j.gie.2020.06.040.
14. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, Feb. 2017, doi: 10.1038/ nature21056.
15. Briganti and O. Le Moine, "Artificial Intelligence in Medicine: Today and Tomorrow," Front Med (Lausanne), vol. 7, Feb. 2020, doi: 10.3389/fmed .2020. 00027.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue III March 2025

16. Kamal Alsheref and W. Hassan Gomaa, "Blood Diseases Detection using Classical Machine Learning Algorithms," 2019. [Online]. Available: www.ijacsa. thesai.org
17. H. Buch, I. Ahmed, and M. Maruthappu, "Artificial intelligence in medicine: Current trends and future possibilities," Mar. 01, 2018, Royal College of General Practitioners. doi: 10.3399/bjgp18X695213.
18. B. Johnson et al., "Precision Medicine, AI, and the Future of Personalized Health Care," Jan. 01, 2021, Blackwell Publishing Ltd. doi: 10.1111/cts.12884.
19. Hamet and J. Tremblay, "Artificial intelligence in medicine," Metabolism, vol. 69, pp. S36–S40, Apr. 2017, doi: 10.1016/j.metabol.2017.01.011.
20. Kaul, S. Enslin, and S. A. Gross, "History of artificial intelligence in medicine," Oct. 01, 2020, Mosby Inc. doi: 10.1016/j.gie.2020.06.040.
21. A. Hoogenboom, U. Bagci, and M. B. Wallace, "Artificial intelligence in gastroenterology. The current state of play and the potential. How will it affect our practice and when?," Apr. 01, 2020, Elsevier B.V. doi: 10.1016/j.tgie.2019.150634.
22. D. S. M. F. M. L. G. L. M. C. G. C. G. C. M. L. H. Jodie C. Avery, "Noninvasive diagnostic imaging for endometriosis part 1: a systematic review of recent developments in ultrasound, combination imaging, and artificial intelligence," Fertil Steril, vol. 121, no. 2, pp. 164–188, Feb. 2024.
23. Ji et al., "U-Net_dc: A Novel U-Net-Based Model for Endometrial Cancer Cell Image Segmentation," Information (Switzerland), vol. 14, no. 7, Jul. 2023, doi: 10.3390/info14070366.
24. Chaddad, J. Peng, J. Xu, and A. Bouridane, "Survey of Explainable AI Techniques in Healthcare," Jan. 01, 2023, MDPI. doi: 10.3390/s23020634..
25. H. Thunold, M. A. Riegler, A. Yazidi, and H. L. Hammer, "A Deep Diagnostic Framework Using Explainable Artificial Intelligence and Clustering," Diagnostics, vol. 13, no. 22, Nov. 2023, doi: 10.3390/diagnostics13223413.