Developing a Fundamental Understanding of Explainable Artificial Intelligence using Texture Analysis

Md Masum Billah[1], Pragati Manandhar[1], Sarosh Krishan[1], Alejandro Cedillo Gamez[1], Kurt K. Benke[2], Sepinoud Azimi[1,3], Janan Arslan[4]

Affiliations: 1. EMJMD Program, EDISS, Faculty of Science and Engineering, Information Technology, Åbo Akademi University, 20500, Turku, Finland.
2. School of Engineering, University of Melbourne, Parkville, Victoria 3010, Australia.
3. Department of Information and Communication Technology, Technology, Policy and Management, Delft University of Technology, Delft, 2600GA, The Netherlands.
4. Sorbonne Université, Institut du Cerveau—Paris Brain Institute—ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013 Paris, France.

Contact Email: Md.Billah@abo.fi

Introduction

As Artificial Intelligence (AI) continues to advance, particularly in healthcare applications like cancer diagnosis, concerns have arisen regarding the opacity of many AI models. The inability to comprehend how these models extract features from medical data for predictions has sparked debates and raised questions about their reliability. In response, Explainable AI (XAI) has emerged as a field dedicated to creating transparent, interpretable, and accountable AI technologies for medical use. While current XAI research utilizes techniques such as Class Activation Maps (CAMs) and tools like SHAP [5] and Quantus [6] to visualize and quantify features influencing AI predictions, these approaches offer only a partial understanding. This project proposes a novel approach to enhance the interpretability of AI models by leveraging texture analysis, particularly in cancer datasets. By focusing on specific texture features within medical images during training, this approach aims to shed light on how AI models make predictions and improve their trustworthiness. Through rigorous analysis of texture features extracted from medical images and correlation with AI predictions, this project aims to bridge the gap between AI predictions and human understanding in cancer diagnosis.

Methodology

The overall pipeline is illustrated in Figure 1.

1. Obtain Trained Segmentation Model: Deep learning models, such as the U-Net and DeepLabv3 were trained on the Curated Breast Imaging Subset of DDSM (CBIS-DDSM) dataset (Figure 2) [7]. This dataset was chosen for its detailed annotations and minimal class imbalance. To enhance model robustness, we employed patch-level segmentation with carefully selected patch sizes, ensuring precise tumor delineation and effective learning outcomes.

2. Extract Texture Features: Advanced methods, including the Gray Level Co-occurrence Matrix (GLCM), Local Binary Patterns (LBP), and Adaptive Texture Energy Measure (aTEM) methods, were utilized to extract texture features from both raw input images and the feature maps generated by each layer of the segmentation models. These extracted texture features were then numerically represented for detailed analysis in subsequent stages of the study.

3. Correlate Texture Energy Map: In the final phase, texture energy maps from each model layer were correlated with the outcomes of texture analysis. Specific layers of interest, such as 'backbone.layer1', 'backbone.layer2', 'backbone.layer3', and 'backbone.layer4', were identified. Correlation coefficients were used to quantify the relationships, ranking emphasized texture types and providing insights into the influential features across different model levels.

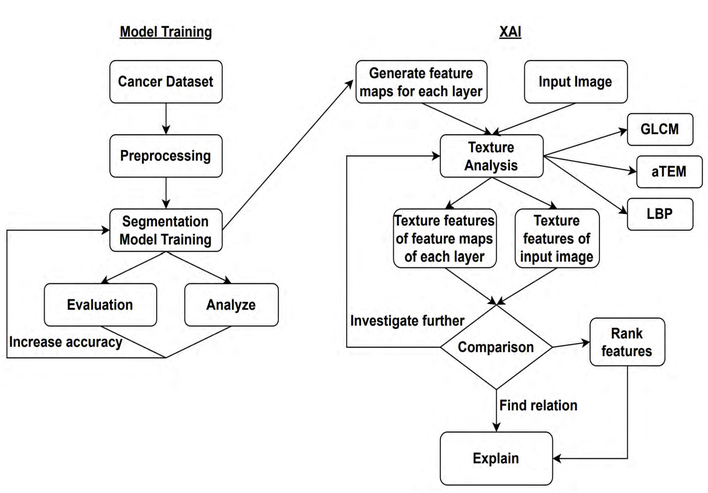


Figure 1: XAI Pipeline.

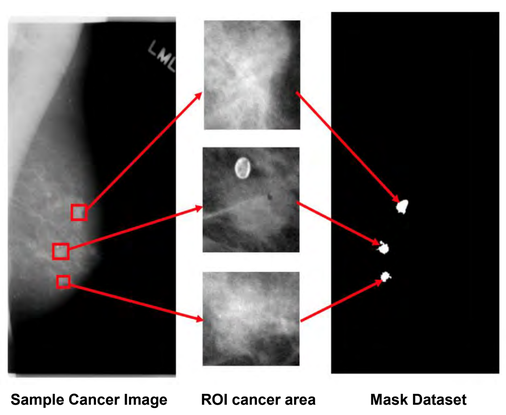


Figure 2: Sample breast cancer image from CBIS-DDSM dataset.

Preliminary Results and Analyses

Two segmentation models, DeepLabv3 and U-Net, were trained, and we obtained mean IoUs of 93.0 and 77.82, respectively. Table 1 shows the models' performance for cancer images.

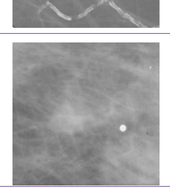
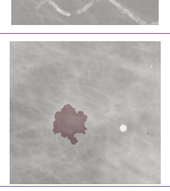
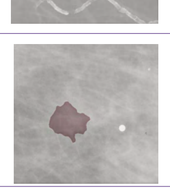
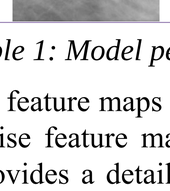
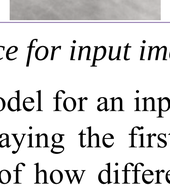
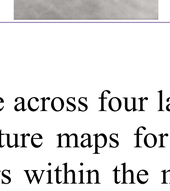
Model	Input Image	Ground Truth	Predicted Mask
DeepLabv3			
U-Net			

Table 1: Model performance for input images

Afterward, we obtained the feature maps of our model for an input image across four layers. Figure 3 illustrates layer-wise feature maps, displaying the first 10 feature maps for each layer. This visualization provides a detailed view of how different layers within the model process and transform the input images, highlighting the evolution of features as they pass through the network.

Then we focused on the texture features that influence the model's predictions of the mask. We extracted GLCM features. There are 13 features extracted from the input image, including Angular Second Moment (ASM), Contrast, Correlation, Variance, Inverse Difference Moment (IDM), Homogeneity, Sum Entropy, Entropy, Difference Entropy, Information Measure of Correlation 1 (IMC1), Information Measure of Correlation 2 (IMC2), Maximal Correlation Coefficient (MCC), and Autocorrelation.

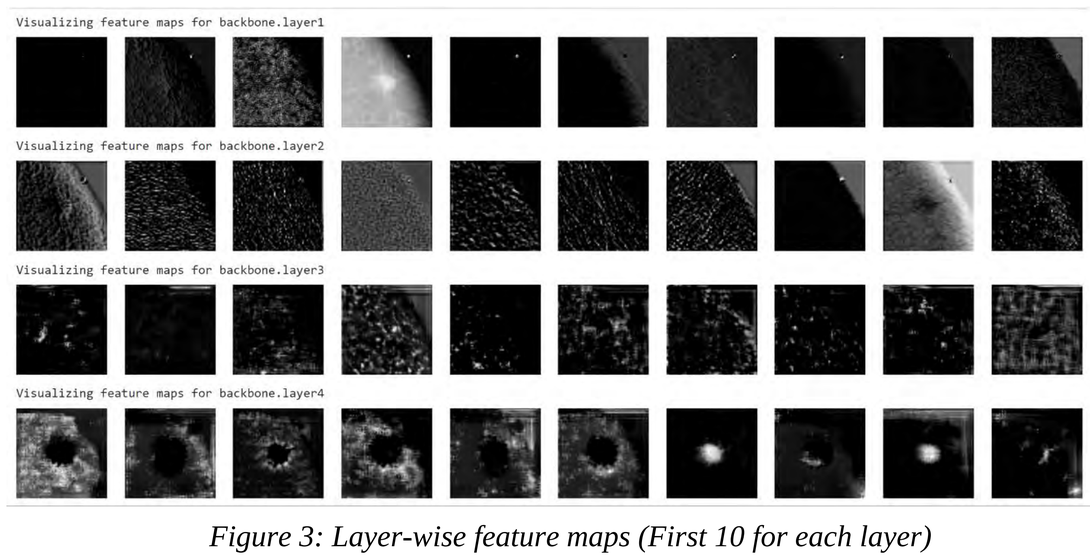


Figure 3: Layer-wise feature maps (First 10 for each layer)

We extracted these 13 GLCM features from 2048 feature maps from each layer (Layer 1, Layer 2, Layer 3, and Layer 4) of the model. Then, we compared the GLCM features of the original image with those of the feature maps using two different approaches.

First Approach: Computes the average absolute differences between the GLCM features of the original image and those of each feature map.

Second Approach: Computes the average distances/similarities between GLCM features of the original image and those of each feature map using Euclidean distance, Manhattan distance, and cosine similarity.

Position	First Approach	Second approach
Layer 1		
First	IMC1 (Score: 0.00116)	IMC1 (Score: 0.33411)
Second	ASM (Score: 0.11807)	ASM (Score: 0.41204)
Third	Autocorrelation (Score: 0.13050)	Autocorrelation (Score: 0.41806)
Layer 2		
First	IMC1 (Score: 0.00084)	IMC1 (Score: 0.33389)
Second	ASM (Score: 0.13291)	ASM (Score: 0.42194)
Third	Autocorrelation (Score: 0.16339)	Autocorrelation (Score: 0.44226)
Layer 3		
First	IMC1 (Score: 0.00265)	IMC1 (Score: 0.33510)
Second	IDM (Score: 0.25426)	IDM (Score: 0.50284)
Third	Autocorrelation (Score: 0.31222)	Autocorrelation (Score: 0.54148)
Layer 4		
First	IMC1 (Score: 0.000755)	IMC1 (Score: 0.3338)
Second	Autocorrelation (Score: 0.02375)	Autocorrelation (Score: 0.3491)
Third	ASM (Score: 0.0745)	ASM (Score: 0.3830)

Table 2: Top three texture features correlated with feature maps.

The analysis reveals the hierarchical influence of texture features within the model layers. Initially, in the first and second layers, IMC1, ASM, and Autocorrelation emerge as pivotal factors. Subsequently, IDM gains prominence in the third layer. Ultimately, Autocorrelation ascends as the second most influential feature in the final layer. Throughout the model's decision-making process, IMC1, Autocorrelation, and ASM consistently exert the most significant influence (Table 2).

Further Works

Our work continues with the successful implementation of the GLCM method in our segmentation model, a crucial step for extracting texture features in cancer diagnosis. Ongoing efforts include integrating the aTEM method to enhance the depth of texture features. Additionally, our correlation method, correlating texture energy maps with analysis outcomes, is in progress. These additions aim to provide a more comprehensive understanding of influential features in different layers, ultimately enhancing the interpretability and accuracy of our AI system for cancer diagnosis.

Conclusion

In summary, our project enhances cancer diagnosis transparency through AI. We trained a robust segmentation model on a relevant dataset, integrated advanced texture analysis, and correlated texture energy maps to reveal influential textures. This ensures more interpretable and trustworthy medical outcomes. Our collaborative approach marks a significant step in bridging the gap between AI predictions and human understanding in cancer diagnosis.

References

1. Laws. (1980). Textured Image Segmentation.
2. Ertugrul. (2014). Adaptive Texture Energy Measure Method. Doi: 10.11648/j.ijiis.20140302.11
3. Ronneberger, Fischer, & Thomas Brox. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597
4. Chen, Papandreou, Schroff, & Adam. (2017). Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv:1706.05587
5. Scott M. Lundberg, Su-In Lee. A Unified Approach to Interpreting Model Predictions. Part of Advances in Neural Information Processing Systems 30 (NIPS 2017).
6. J. Smith et al. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond . arXiv:2202.06861.
7. Lee, R., Gimenez, F., Hoogi, A. et al. A curated mammography data set for use in computer-aided detection and diagnosis research. Sci Data 4, 170177 (2017). https://doi.org/10.1038/sdata.2017.177

Funded by the European Union

Erasmus Mundus Association (EMA)
2nd International Poster Competition 2024

Erasmus Mundus Association