European Journal of Radiology 173 (2024) 111356

Contents lists available at ScienceDirect

European Journal of Radiology

journal homepage: www.elsevier.com/locate/ejrad

Research article

Quantitative evaluation of Saliency-Based Explainable artificial intelligence (XAI) methods in Deep Learning-Based mammogram analysis

Esma Cerekci [a], Deniz Alis [b,*], Nurper Denizoglu [c], Ozden Camurdan [c], Mustafa Ege Seker [d], Caner Ozer [e], Muhammed Yusuf Hansu [f], Toygar Tanyel [g], Ilkay Oksuz [e], Ercan Karaarslan [b]

[a] Sisli Hamidiye Etfal Training and Research Hospital, Department of Radiology, Istanbul, Turkey
[b] Acibadem Mehmet Ali Aydinlar University, School of Medicine, Department of Radiology, Istanbul, Turkey
[c] Acibadem Healthcare Group, Department of Radiology, Istanbul, Turkey
[d] Acibadem Mehmet Ali Aydinlar University, School of Medicine, Istanbul, Turkey
[e] Istanbul Technical University, Department of Computer Engineering, Istanbul, Turkey
[f] Istanbul Technical University, Department of Electronics and Communication Engineering, Istanbul, Turkey
[g] Istanbul Technical University, Department of Biomedical Engineering, Istanbul, Turkey

ARTICLE INFO

Keywords:
Breast Cancer
Deep Learning
Mammogram
XAI

ABSTRACT

Background: Explainable Artificial Intelligence (XAI) is prominent in the diagnostics of opaque deep learning (DL) models, especially in medical imaging. Saliency methods are commonly used, yet there's a lack of quantitative evidence regarding their performance.
Objectives: To quantitatively evaluate the performance of widely utilized saliency XAI methods in the task of breast cancer detection on mammograms.
Methods: Three radiologists drew ground-truth boxes on a balanced mammogram dataset of women (n = 1496 cancer-positive and negative scans) from three centers. A modified, pre-trained DL model was employed for breast cancer detection, using MLO and CC images. Saliency XAI methods, including Gradient-weighted Class Activation Mapping (Grad-CAM), Grad-CAM++, and Eigen-CAM, were evaluated. We utilized the Pointing Game to assess these methods, determining if the maximum value of a saliency map aligned with the bounding boxes, representing the ratio of correctly identified lesions among all cancer patients, with a value ranging from 0 to 1.
Results: The development sample included 2,244 women (75%), with the remaining 748 women (25%) in the testing set for unbiased XAI evaluation. The model's recall, precision, accuracy, and F1-Score in identifying cancer in the testing set were 69%, 88%, 80%, and 0.77, respectively. The Pointing Game Scores for Grad-CAM, Grad-CAM++, and Eigen-CAM were 0.41, 0.30, and 0.35 in women with cancer and marginally increased to 0.41, 0.31, and 0.36 when considering only true-positive samples.
Conclusions: While saliency-based methods provide some degree of explainability, they frequently fall short in delineating how DL models arrive at decisions in a considerable number of instances.

1. Introduction

Mammographic examinations stand at the forefront of breast cancer detection due to their widespread availability, cost-effectiveness, and reliable sensitivity in identifying anomalies [1,2]. However, the increasing demand for mammogram interpretation imposes a significant burden on radiologists, leading to delayed reports, overlooked examinations, and a potential increase in diagnostic errors [3–6].

In recent years, advancements in hardware and software, coupled with the exponential growth of digital medical data, have positioned

Abbreviations: **BI-RADS**, Breast Imaging-Reporting and Data System; **CC**, Cranio-caudal; **CNN**, Convolutional Neural Network; **DICOM**, Digital Imaging and Communications in Medicine; **DL**, Deep learning; **Grad-CAM**, Gradient-weighted Class Activation Mapping; **MLO**, Mediolateral-oblique; **XAI**, Explainable artificial intelligence.

* Corresponding author at: Acibadem Mehmet Ali Aydinlar University, School of Medicine, Department of Radiology, Atasehir/ Istanbul, Turkey.
E-mail addresses: esmaktufan@gmail.com (E. Cerekci), drdenizalis@gmail.com (D. Alis), nurperonder@hotmail.com (N. Denizoglu), ozlemcamurdan@gmail.com (O. Camurdan), smustafaege@gmail.com (M. Ege Seker), ozlemcamurdan@gmail.com (C. Ozer), myusufhansu@gmail.com (M.Y. Hansu), tanyel23@itu.edu.tr (T. Tanyel), oksuzilkay@itu.edu.tr (I. Oksuz).

https://doi.org/10.1016/j.ejrad.2024.111356
Received 23 August 2023; Received in revised form 10 December 2023; Accepted 2 February 2024
Available online 5 February 2024

deep learning (DL) as a promising solution. DL offers assistance to medical professionals in diagnosing a spectrum of diseases, handling a variety of tasks [7–10]. Mammography assessment is no exception to this trend, with numerous studies documenting DL's ability to match or even exceed the performance of radiologists in identifying breast cancer in mammograms [11,12].

DL models can identify the most relevant features and perform downstream tasks (such as classification, segmentation, detection) without relying on hand-crafted features [13]. Despite their effectiveness, these models, often containing millions or even billions of parameters, are referred to as "black-box" or "opaque" due to their inherent complexity and limited transparency in their decision-making processes [13]. This opaque nature of DL models has raised concerns in medical applications where decisions typically involve high-stake tasks, such as breast cancer detection.

To address the 'black-box' nature of DL models, a diverse range of methods, collectively known as explainable artificial intelligence (XAI), have been proposed. XAI comprises a set of techniques designed to improve the transparency and interpretability of DL, encompassing areas such as algorithmic transparency, uncertainty analysis, and model visualization and inspection [14].

A recent systematic review on XAI in medical imaging indicates that approximately one-third of studies implemented a type of XAI method, with visualization or saliency-based techniques being the most commonly utilized to understand a model's inner workings [15]. Nevertheless, most of these studies evaluated XAI methods qualitatively, often presenting selectively chosen examples [15,16]. A limited number of studies have quantitatively investigated the performance of XAI methods [17–19], underscoring the need for further research evaluating the efficacy of commonly applied XAI methods.

In this study, we quantitatively evaluate the performance of widely utilized saliency-based XAI methods—Gradient-weighted Class Activation Mapping (Grad-CAM) [20], Grad-CAM++ [21], and Eigen-CAM [22]—in the task of breast cancer detection on mammograms using the Pointing Game Score [23].

2. Methods

2.1. Study sample

This retrospective study, approved by the local review board, which also waived the requirement for informed consent for the retrospective review and analysis of anonymized medical data, utilized data from our university hospitals (n = 3). We searched the hospital information system and picture archive and communication system between January 2016 and January 2020 to identify women aged > 18 years who had undergone a screening mammogram (n = 28324). Patients with prior history of breast surgeries, radiotherapy or chemotherapy treatment before their screening mammogram were excluded from the study (n = 124).

Subsequently, we identified women with pathology-proven breast cancer following their screening and included them in the breast cancer group (n = 1496). To provide a more comprehensive and realistic representation of clinical cases, our study included all types of breast cancer lesions, whether presenting as a mass, microcalcification, or a combination of both.

To establish a non-breast-cancer group, we included women with a Breast Imaging-Reporting and Data System (BI-RADS) [24] score of 1 or 2 on their mammogram, with at least one follow-up screening mammogram obtained one or more years after the initial mammogram, or women with pathology-proven benign findings. To create a balanced dataset for our experiments, we under-sampled the non-breast-cancer group to match the size of the breast-cancer group (n = 1496). This approach, while potentially introducing bias, was necessary to prevent class imbalance.

Any potential biases introduced due to under-sampling were mitigated by using stratified random sampling to ensure that our non-breast-cancer group remained representative of the initial population of women. This process ensured that important subgroups (e.g., age groups, BI-RADS scores, etc.) were proportionally represented.

Lastly, the initial study sample was divided into a development set (comprising of the training and validation data) and a held-out testing set using a 75 %/25 % split, stratified to ensure representation of all classes.

2.2. Ground truthing

Ground truthing was conducted by three board-certified breast radiologists working in consensus, each with a minimum of five years of specialized experience in breast imaging.

Initially, the team rigorously reviewed the mammography images from the non-breast-cancer group, cross-referencing these images with the clinical and pathology reports to confirm them as true negative cases. Once validated, the radiologists proceeded to perform annotations for mammography examinations within the breast-cancer group, on a dedicated workstation equipped with a dedicated browser-based platform (https://matrix.md.ai) with a 6-megapixel diagnostic monitor (Radiforce RX 660, EIZO). All reviewed images were in Digital Imaging and Communications in Medicine (DICOM) format.

The radiologists were tasked with drawing bounding boxes around lesions confirmed to harbor breast cancer. This task was performed with the guidance of pathology reports and clinical notes, serving as reliable references. In instances where a mammogram revealed two distinct foci of breast cancer, the radiologists were instructed to denote these separate findings with two distinct bounding boxes. These bounding boxes were consistently placed on both cranio-caudal (CC) and mediolateral-oblique (MLO) views for each patient, ensuring a comprehensive representation of the breast cancer location and extent.

We did not seek independent ground-truthing process as assessing the inter-rater agreement was not within the scope of this study.

2.3. DL model

All DL experiments in this study were carried out utilizing the PyTorch Library Version 1.7.1. We selected the ResNet50 model [25], pre-trained on ImageNet, due to its established prevalence and robust performance within the medical imaging domain, albeit with several modifications to adapt it to our specific use case.

The mammograms in our dataset were processed using a custom-developed algorithm designed to minimize empty spaces surrounding the breast tissue. By employing a patch-based analysis approach, the algorithm identifies the largest rectangle that can encompass each breast, facilitating image cropping without loss of critical information from the breast area. Following this, the images were padded and resampled to maintain a uniform resolution of 512x512 pixels. Prior to their integration into the model, all images underwent a review by a radiologist to confirm the presence of all necessary information and the effectiveness of the cropping function.

The model trained as a multi-view classification network, feeding the MLO and CC images of the same breast via 2 input channels. This allowed model to capture correlated but complementary information from both views and has been consistently shown to increase the performance compared with the uni-view DL models in earlier studies [26].

In a critical adaptation, the original fully connected layer was replaced with a layer that combines a linear transformation with a sigmoid activation function. This alteration enabled us to perform binary classification and derive probabilities for each class, enhancing our model's interpretability.

The Cross-Entropy Loss function was employed as our loss function and the Adam optimizer was utilized to train the model, selected for their effectiveness in training classification models. Following careful consideration of computing efficiency and training time, we have set the

batch size to 16. The epoch number was configured to 100 which provides model sufficient iterations to learn from dataset while mitigating the risk of overfitting.

To expedite training efficiency and promote better convergence, we implemented a learning rate scheduler. Specifically, we opted for the StepLR scheduler, which progressively decreases the learning rate by a factor of 0.5 every 20 epochs, commencing with a learning rate of 0.0001. This strategy was adopted to allow the model to refine its parameters more adeptly during the latter stages of training, thereby guiding it to a potentially more optimal solution. Standard data augmentation methods such as transpose, vertical and horizontal flip, random shifts, scales, and rotations were used on the fly. Fig. 1 shows the DL pipeline used in the present work.

2.4. XAI methods

We used Grad-CAM) [20], Grad-CAM++ [21] and Eigen-CAM [22] as XAI methods in this study. The following paragraphs briefly explain each method without delving into too much technical details.

Grad-CAM is a widely-used explanation method for Convolutional Neural Networks (CNNs) [20]. It generates "visual explanations" for decisions from a wide variety of CNN-based models. It utilizes the gradient information flowing into the final convolutional layer of the CNN to understand each neuron for a decision of interest. The generated heatmaps, known as Grad-CAMs, highlight regions in the input that were influential for the network's output, providing insight into the model's decision-making process.

Grad-CAM++ is an extension of Grad-CAM [21]. This method considers multi-layer and multi-node information for visual explanations, unlike Grad-CAM, which focuses on the final convolutional layer. The highlight of Grad-CAM++ is its capability to capture multi-modal behavior in the feature maps, as opposed to just peak behavior. Consequently, Grad-CAM++ may provide more detailed visual explanations and can better address the challenge of identifying multiple regions of interest in the input.

Eigen-CAM is another method for understanding the decision-making process of DL models, especially CNNs [22]. Eigen-CAM leverages eigen-decomposition of the spatial dimensions of the final convolution layer's feature maps. By combining the corresponding eigenvalues and eigenvectors, it generates the class activation maps. Unlike Grad-CAM and Grad-CAM++, Eigen-CAM can handle negative values in the feature maps, making it more suitable for explaining certain models, such as those with Batch Normalization layers or ReLU layers.

2.5. XAI evaluation metrics

In our study, the Pointing Game [23] was employed as a quantitative measure to evaluate the performance of XAI methods in aligning with clinical ground truths. The primary focus was to ascertain the correlation between the saliency maps produced by these methods and the ground-truth bounding boxes, which were delineated by radiologists on mammogram lesions suspected of harboring breast cancer.

The Pointing Game's core assessment involves determining whether the saliency map's maximum value falls within the ground-truth bounding boxes, factoring in an offset denoted as τ. In our methodology, we adhered to a default offset value of $\tau = 15$. This value was chosen based on input from breast cancer experts, as it offers a realistic approximation of the model's focus areas in relation to the clinically identified lesions.

In our study, we adopted an image-level analysis for evaluating the Pointing Game score due to the unique nature of breast imaging. Lesions in mammography may be more pronounced or solely visible in one view (MLO or CC), influenced by factors like breast composition and lesion characteristics. Independently assessing each view allows for a more detailed and accurate evaluation of saliency methods, catering to the variability in lesion detection across different views. Consequently, when a lesion was distinctly highlighted in one view but not the other, each view was evaluated separately for the Pointing Game score, ensuring a thorough and nuanced assessment of the saliency methods' performance in mammographic analysis.

To calculate the Pointing Game score, the process involved identifying the pixel with the highest activation level within each saliency map, and then evaluating if this maximum value was confined within the expert-drawn bounding boxes, accounting for the $\tau = 15$ offset. The rationale behind this approach and its implications are further elucidated in Fig. 2, which visually demonstrates the Pointing Game metric with varying τ values, thus providing a comprehensive understanding of the assessment methodology.

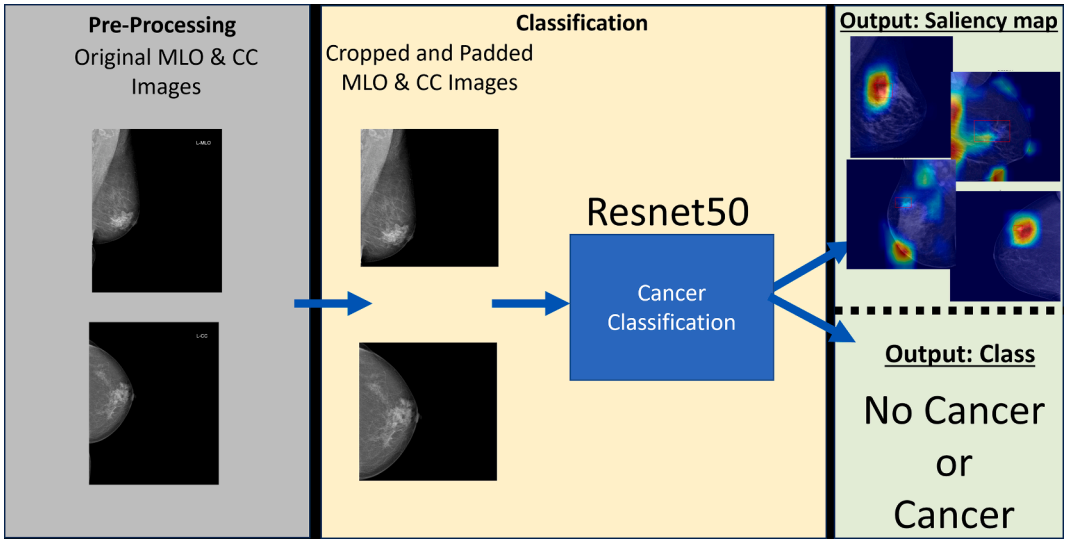

Fig. 1. The Deep Learning Model Used in This Study. The mammograms were cropped to minimize empty spaces surrounding the breast tissue and padded to maintain a uniform resolution of 512x512 pixels. We trained a 2-channel Resnet-50 architecture to detect breast cancer. We utilize saliency map techniques—Gradient-weighted Class Activation Mapping (Grad-CAM), Grad-CAM++, and Eigen-CAM—to highlight the regions where the classification model focused on while making its decisions.

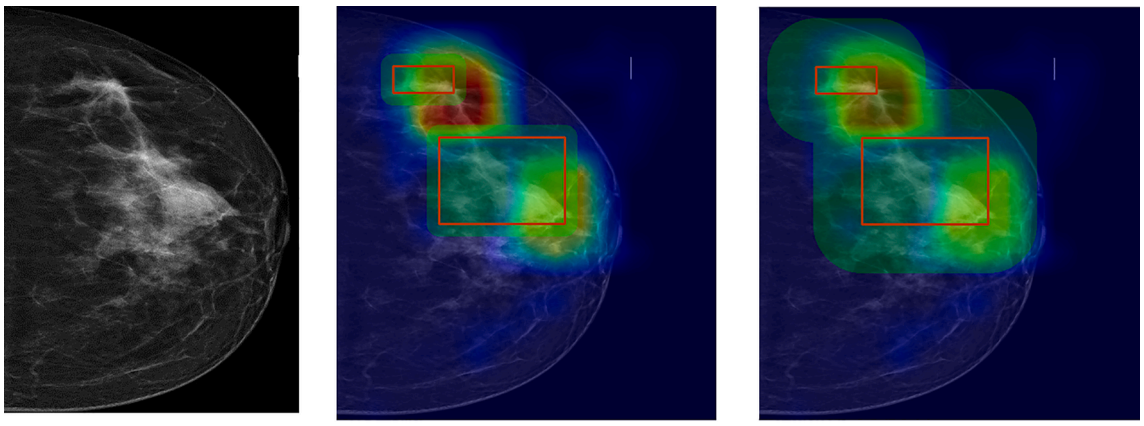

Fig. 2. Pointing Game metric with different τ values. We illustrate the influence of τ parameter on pointing game score. Original image is shown in Fig. 2a. Pointing game aims to evaluate the saliency map methods by focusing on a neighborhood around the peak value. We showcase the influence of varying τ (15,60) on an example case. The τ parameter is adjustable and gives a certain degree of offset with respect to the ground-truth boxes in calculating whether the peak value of the saliency map lies within the ground-truth bounding boxes. In this example image, the green rounded rectangle around the ground-truth boxes represents the area where if the peak value of the saliency map overlaps, it is accepted as hit. As the τ increases, this area increases, and the chance of a hit also increases. We used the default τ of 15 in this study. The ground-truth boxes (a), τ of 15 (b), and 60 (c) are shown.

2.6. Statistical analyses

Statistical analyses were conducted using Python Version 3′s SciPy library. As the data were not normally distributed, continuous variables are presented using median values and interquartile ranges, while categorical and ordinal variables are illustrated with frequencies and percentages. The DL model's performance in identifying breast cancer is measured using recall, precision, accuracy, and the F1-score. For the accuracy, recall, and precision, we calculated 95 % confidence intervals using the method following prior work [27].

We assessed the performance of saliency maps using the Pointing Game Score as follows: If the maximum intensity point on the saliency map lies within close proximity to the ground truth bounding boxes, we categorize the saliency map as accurate. Conversely, if this point lies outside the defined range, we consider the saliency map inaccurate. Readers should bear in mind that the Pointing Game Score were calculated only for patients with cancer as the model in this study trained to detect cancer lesions while treating patients with benign lesions or negative mammograms (BI-RADS 1–2) as negative cases (i.e., no bounding boxes are provided). The number of accurate and inaccurate saliency maps are denoted as T and F respectively, allowing us to formulate the accuracy (i.e., the Pointing Game Score) as follows [23]:

$$A = T/(T+F)$$

We compared the Pointing Game Scores of Grad-CAM, Grad-CAM++, and Eigen-CAM using the Wilcoxon test, with a p-value of less than 0.05 considered indicative of a statistically significant result.

3. Results

A total of 2,992 women, with a median age of 49 (IQR, 8), constituted the final study sample. Among these women, 1,496 (50 %) had pathology-proven breast cancer, while the remaining 1,496 (50 %) had negative mammograms. Only a single mammography scan included per patient in the present work. The development sample included 2,244 women (75 % of the total sample), while the held-out testing set, utilized for the unbiased evaluation of the XAI methods, comprised the remaining 748 women (25 %).

The recall, precision, accuracy, and F1-Score of the DL model in identifying breasts with cancer in the testing set were 69 % (95 % CI, 0.621–0.759), 88 % (95 % CI, 0.828–0.932), 80 % (95 % CI, 0.735–0.865), and 0.77, respectively.

In the testing set, the Pointing Game Scores for Grad-CAM, Grad-CAM++, and Eigen-CAM were 0.41, 0.30, and 0.35, respectively in women with cancer. No statistically significant difference was observed between each method as per the Kruskal-Wallis test. However, when only true-positive samples were considered, the Pointing Game Scores for Grad-CAM, Grad-CAM++, and Eigen-CAM increased marginally to 0.41, 0.31, and 0.36, respectively.

Figs. 3, 4, and 5 display representative examples of the XAI methods from the testing set.

4. Discussion

DL applications for medical imaging in diagnosis have attracted significant attention in recent years, with many emerging academic and clinical applications. Though DL's success is well-recognized, the black-box nature of DL methods has raised concerns. XAI methods aim to address the opacity of DL methods, yet most studies implementing XAI methods for DL in medical imaging for diagnosis have only qualitatively evaluated the benefits of XAI and have presented cherry-picked examples without systematic analysis regarding their performance.

In this study, we conducted a quantitative evaluation of the performance of a collection of widely utilized XAI methods, specifically saliency-based techniques including Grad-CAM [20], Grad-CAM++ [21], and Eigen-CAM [22], within the domain of medical imaging. These methods were applied to one of the most prevalent uses of DL in medical imaging: breast cancer detection on mammography. For our analysis, we employed a Pointing Game Score, using bounding boxes determined by experts as the ground truth.

The most conspicuous result arising from this study is the dichotomy in performance. While these methods exhibited substantial success in certain cases, their quantitative performance in accurately pinpointing the location of pathology (i.e., elucidating the model's decision) was relatively low. They failed in a significant number of other instances. Furthermore, our investigation revealed that, despite a slight enhancement in performance, these methods yielded only slight increase Pointing Game Score scores in situations where the model correctly identified the presence of cancer. In comparing the three approaches, Grad-CAM emerged as the method with the highest performance

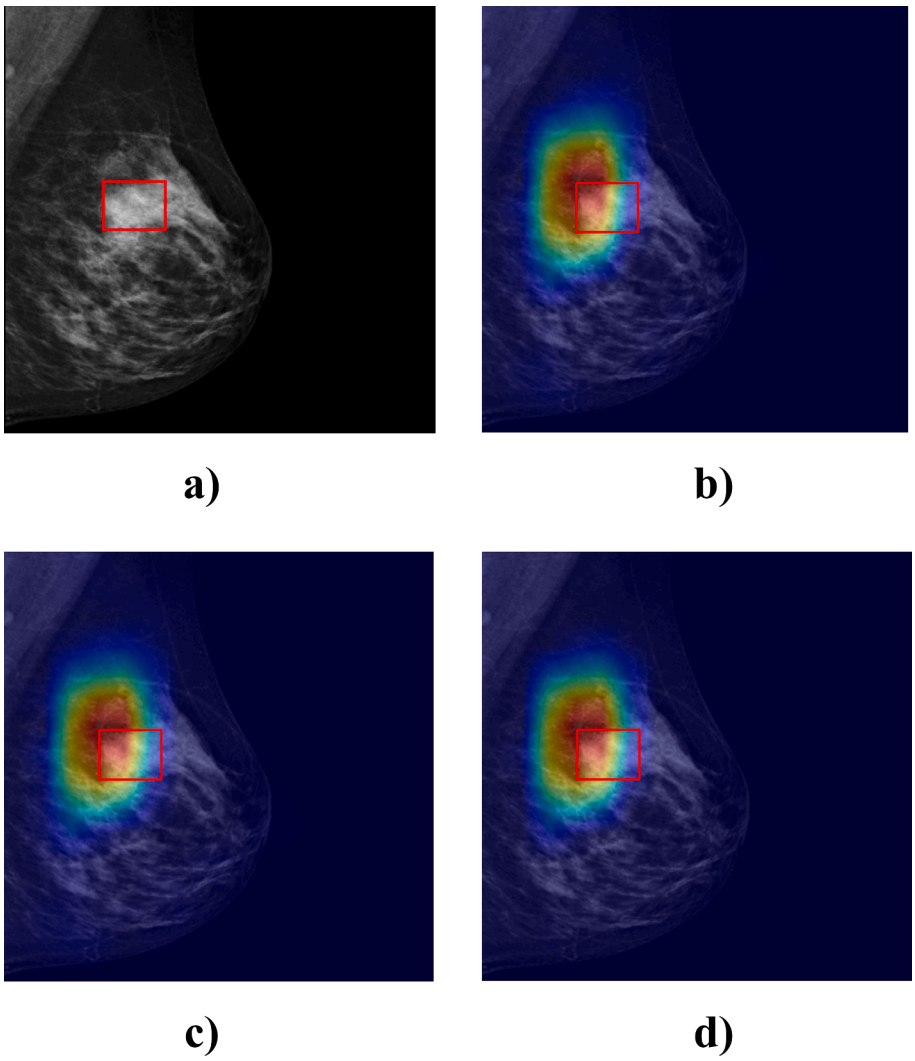

Fig. 3. Original image with the ground-truth box (a), Gradient-weighted Class Activation Mapping (Grad-CAM) (b), Grad-CAM++ (c), and Eigen-CAM (d) maps in a woman in whom the DL model correctly classified breast cancer. All saliency methods accurately pinpoint the lesion (i.e., considered accurate by Pointing Game Metric) using the ground-truth box as the reference (rectangle).

amongst all.

Our findings are consistent with those from the study by Arun et al., in which the authors meticulously investigated the performance of various saliency-based techniques in the task of detecting pneumothorax and pneumonia on publicly available datasets [18]. They employed different methodologies to scrutinize aspects of saliency-based methods such as utility, repeatability, and localization performance, employing measures akin to those in our study. However, they assessed localization performance using the area under the precision-recall curve. A direct comparison with our results is not feasible, as the authors investigated XAI methods across different imaging modalities and utilized a variety of distinct metrics. Nonetheless, they too discovered that the saliency maps underperformed quantitatively compared to the average of bounding boxes from their dataset, highlighting the limitations of these techniques.

Wollek et al. explored the performance of saliency-based and attention-based methods, employing vision transformers in pneumothorax classification across several publicly available chest X-ray datasets [19]. They evaluated XAI methods using an array of metrics, including positive/negative perturbation, sensitivity-n, effective heat ratio, intra-architecture repeatability, and inter-architecture reproducibility, and assessed the utility of XAI methods in diagnostic decision-making by radiologists. Their findings indicate that radiologists found attention-based XAI methods more beneficial in decision-making compared to Grad-Cam. Again, a direct comparison with our results is not applicable due to the difference in imaging modalities and metrics. However, notably, the effective heat ratio—a metric used in their study to assess the overlap between the saliency map's binary mask and the ground-truth bounding boxes—ranged from 0.16 to 0.33 across different methods in their research. The authors concluded by emphasizing the superiority of attention-based methods over Grad-Cam but did not extensively investigate the circumstances where XAI methods failed.

In this work, the employed ResNet50 model trained on a development set comprising 2,244 women achieved an accuracy of 88 %. This was accomplished using only image-level annotations and a relatively straightforward application of the model. Comparable studies utilizing small-scale datasets and conventional CNN architectures have reported similar performance in mammography-based breast cancer detection [28,29].

Notwithstanding, advancements in employing large-scale datasets, pixel-level annotations, and sophisticated models like vision transformers and next-generation CNNs, have shown potential in significantly enhancing diagnostic performance, potentially surpassing that of radiologists [30,31].

Furthermore, to assess the influence of the model's performance on the effectiveness of saliency maps, we calculated the Pointing Game

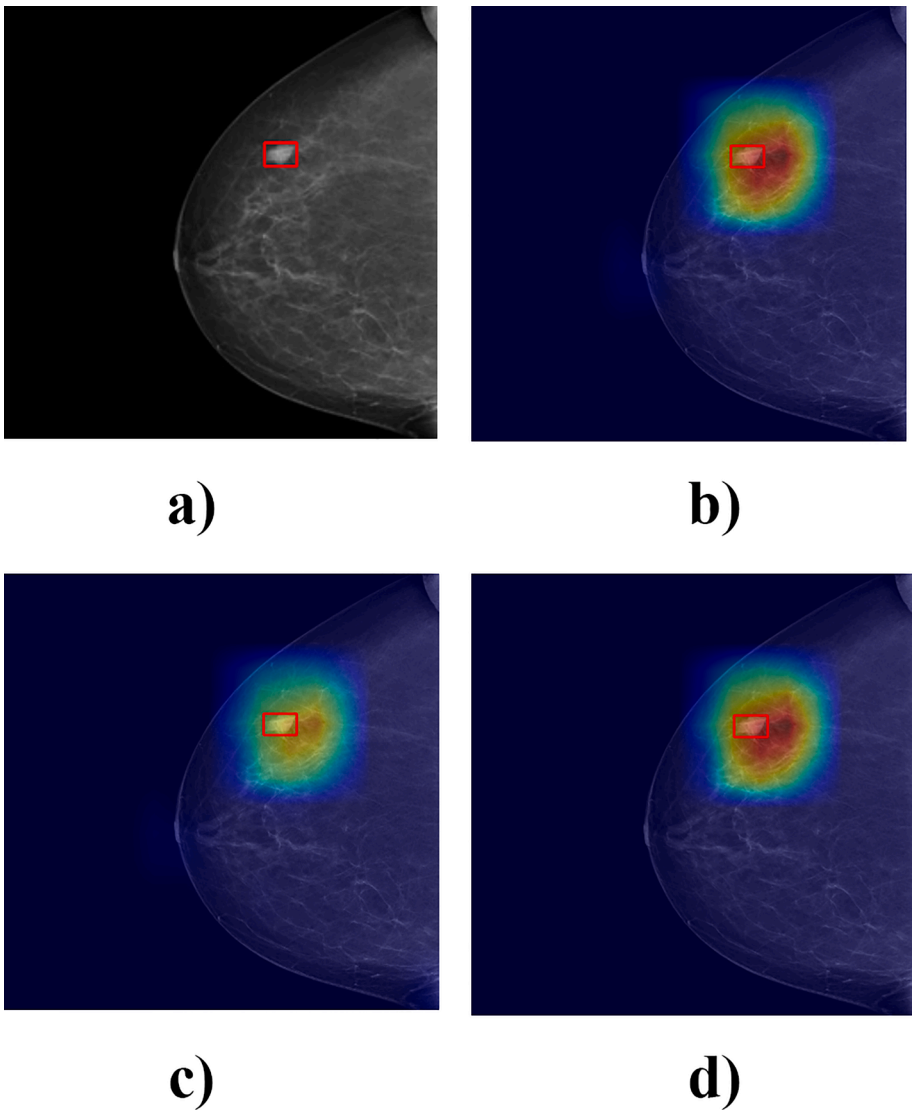

Fig. 4. Original image with the ground-truth box (a), Gradient-weighted Class Activation Mapping (Grad-CAM) (b), Grad-CAM++ (c), and Eigen-CAM (d) maps in a woman in whom the DL model correctly classified breast cancer. Grad-CAM and Eigen-CAM are capable of indicating correct region, where Grad-CAM ++ fails to highlight the lesion area.

score for true-positive samples. This analysis revealed only a marginal increase in the score, underscoring our observation that saliency-based methods often fall short in precisely delineating the decision-making processes of DL models.

In interpreting our study's results, several limitations should be considered. First, the sample size employed in our study was relatively small, potentially affecting the generalizability of our findings.

Second, the study was confined to specific saliency-based methods and did not encompass the entire range of available techniques within the domain. This choice may have led to an incomplete picture of the performance landscape for XAI in medical imaging. In a similar vein, our analysis was limited to CNNs and did not extend to vision transformers or attention-based methods. This restriction could limit the applicability of our findings to the broader range of emerging methods in DL for medical imaging.

Third, it is important for readers to understand that saliency-based XAI methods are primarily focused on highlighting areas of attention in an image as determined by the model. These methods do not offer a comprehensive exploration of the model's inner decision-making processes in the same way that more in-depth methods like Network Dissection or visualizing concept activations do [32,33]. However, it should also be noted that these high-fidelity methods come with their own set of interpretability challenges and require substantial time and effort to analyze.

Fourth, although our downsampling to a resolution of 512x512 is higher compared to the standard 224x224, it is still a reduction from the original resolution typical in mammography. This downsampling might have affected the quality of the saliency maps and subsequently the performance evaluation of the XAI methods.

In conclusion, while saliency-based methods offer explainability to a certain degree, they often fall short in precisely delineating how DL models arrive at a decision in numerous instances. Given that saliency-based techniques are the prevailing choice for XAI in both clinical and academic applications, this underscores the necessity for researchers and clinicians alike to approach them with a clear understanding of their limitations.

Declarations.

Ethical statement.

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

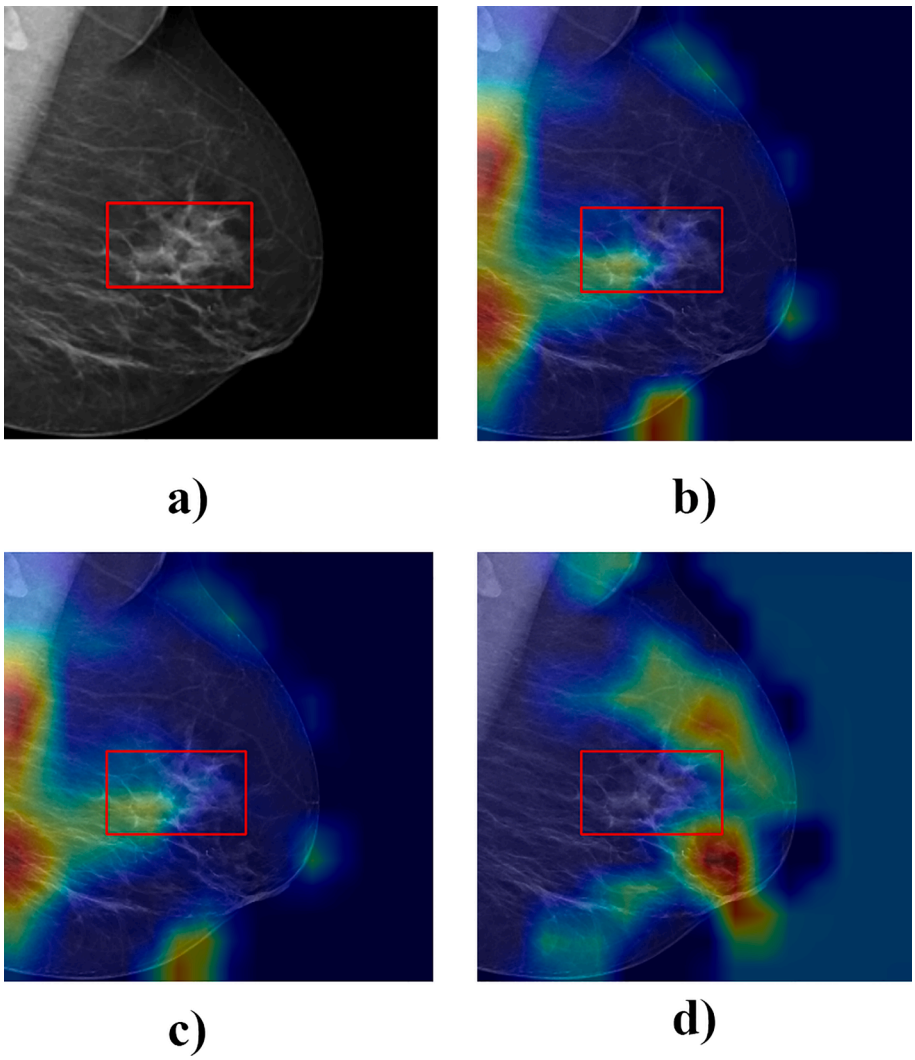

Fig. 5. Original image with the ground-truth box (a), Gradient-weighted Class Activation Mapping (Grad-CAM) (b), Grad-CAM++ (c), and Eigen-CAM (d) maps in a woman in whom the DL model correctly classified breast cancer. All three methods showcase their peak in a region, where no lesion is present. Even though Grad-CAM and Grad-CAM ++ show higher values within the lesion box, they are not capable to focus on the lesion area fully.

The local ethics committee approved this retrospective study and waived the need for informed consent for the retrospective evaluation of anonymized medical data (Acıbadem University and Acıbadem Healthcare Institutions Medical Research Ethics Committee).

Consent for publication

This retrospective study, approved by the local review board, which also waived the requirement for informed consent for the retrospective review and analysis of anonymized medical data, utilized data from our university hospitals.

Funding.

This research did not receive any specific grant from funding agencies in the public, commercial,or not-for-profit sectors.

CRediT authorship contribution statement

Esma Aktufan Cerekci: Writing – review & editing, Writing – original draft, Data curation. **Deniz Alis:** Writing – review & editing, Writing – original draft, Conceptualization. **Nurper Denizoglu:** Data curation. **Ozden Camurdan:** Data curation. **Mustafa Ege Seker:** Formal analysis. **Caner Ozer:** Formal analysis. **Muhammed Yusuf Hansu:** Formal analysis. **Toygar Tanyel:** Formal analysis. **Ilkay Oksuz:** Writing – review & editing. **Ercan Karaarslan:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This paper has been published benefiting from the 1001 Science and Technology National Grant Program of TUBITAK (Project no: 122E022). However, the entire responsibility of the paper belongs to the owner of the paper. The financial support received from TUBITAK does not mean that the content of the publciation is approved in a scientific sense by TUBITAK.

References

[1] K.C. Oeffinger, E.T. Fontham, R. Etzioni, A. Herzig, J.S. Michaelson, Y.-C.-T. Shih, L.C. Walter, T.R. Church, C.R. Flowers, S.J. LaMonte, Breast cancer screening for women at average risk:2015 guideline update from the American Cancer Society, JAMA. 314 (2015) (2015) 1599–1614.

[2] A.L. Siu, U.P.S.T. Force, Screening for breast cancer: US Preventive Services Task Force recommendation statement, Ann. Intern. Med. 164 (2016) 279–296.

[3] J.G. Elmore, S.L. Jackson, L. Abraham, D.L. Miglioretti, P.A. Carney, B.M. Geller, B. C. Yankaskas, K. Kerlikowske, T. Onega, R.D. Rosenberg, Variability in interpretive

performance at screening mammography and radiologists' characteristics associated with accuracy, Radiology. 253 (2009) 641–651.
[4] C.D. Lehman, R.D. Wellman, D.S. Buist, K. Kerlikowske, A.N. Tosteson, D. L. Miglioretti, B.C.S. Consortium, Diagnostic accuracy of digital screening mammography with and without computer-aided detection, JAMA Intern. Med. 175 (2015) 1828–1837.
[5] A.N. Tosteson, D.G. Fryback, C.S. Hammond, L.G. Hanna, M.R. Grove, M. Brown, Q. Wang, K. Lindfors, E.D. Pisano, Consequences of false-positive screening mammograms, JAMA Intern. Med. 174 (2014) 954–961.
[6] N. Houssami, K. Hunter, The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening, npj Breast Cancer. 3 (2017) 12.
[7] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in, Proc. IEEE Conf. Comput. vis. Pattern Recognit. (2017) 2097–2106.
[8] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature. 542 (2017) 115–118.
[9] J. De Fauw, J.R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, Clinically applicable deep learning for diagnosis and referral in retinal disease, Nat. Med. 24 (2018) 1342–1350.
[10] D. Ardila, A.P. Kiraly, S. Bharadwaj, B. Choi, J.J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography, Nat. Med. 25 (2019) 954–961.
[11] K.J. Geras, R.M. Mann, L. Moy, Artificial intelligence for mammography and digital breast tomosynthesis: current concepts and future perspectives, Radiology. 293 (2019) 246–259.
[12] A. Rodriguez-Ruiz, K. Lång, A. Gubern-Merida, M. Broeders, G. Gennaro, P. Clauser, T.H. Helbich, M. Chevalier, T. Tan, T. Mertelmeier, Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists, JNCI J. Natl. Cancer Inst. 111 (2019) 916–922.
[13] G. Chartrand, P.M. Cheng, E. Vorontsov, M. Drozdzal, S. Turcotte, C.J. Pal, S. Kadoury, A. Tang, Deep learning: a primer for radiologists, Radiographics. 37 (2017) 2113–2131.
[14] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion. 58 (2020) 82–115.
[15] A.M. Groen, R. Kraan, S.F. Amirkhan, J.G. Daams, M. Maas, A systematic review on the use of explainability in deep learning systems for computer aided diagnosis in radiology: limited use of explainable AI? Eur. J. Radiol. (2022) 110592.
[16] S. Nazir, D.M. Dickson, M.U. Akram, Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks, Comput. Biol. Med. (2023) 106668.
[17] F. Dong, R. She, C. Cui, S. Shi, X. Hu, J. Zeng, H. Wu, J. Xu, Y. Zhang, One step further into the blackbox: a pilot study of how to build more confidence around an AI-based decision system of breast nodule assessment in 2D ultrasound, Eur. Radiol. 31 (2021) 4991–5000.
[18] N. Arun, N. Gaw, P. Singh, K. Chang, M. Aggarwal, B. Chen, K. Hoebel, S. Gupta, J. Patel, M. Gidwani, Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging, Radiol. Artif. Intell. 3 (2021) e200267.
[19] A. Wollek, R. Graf, S. Čečatka, N. Fink, T. Willem, B.O. Sabel, T. Lasser, Attention-based saliency maps improve interpretability of pneumothorax classification, Radiol. Artif. Intell. 5 (2022) e220187.
[20] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proc. IEEE Int. Conf. Comput. Vis. (2017) 618–626.
[21] A. Chattopadhay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-cam++: generalized gradient-based visual explanations for deep convolutional networks, 2018 IEEE Winter Conf. Appl. Comput Vis. WACV, IEEE. (2018) 839–847.
[22] M. Bany Muhammad, M. Yeasin, Eigen-CAM: visual explanations for deep convolutional neural networks, SN Comput. Sci. 2 (2021) 1–14.
[23] C. Ozer, I. Oksuz, Explainable image quality analysis of chest x-rays, in, Med. Imaging Deep Learn. (2021).
[24] C.J. D'Orsi, E. Sickles, E. Mendelson, E. Morris, ACR BI-RADS atlas: breast imaging reporting and data system; mammography, ultrasound, magnetic resonance imaging, follow-up and outcome monitoring, data dictionary, ACR Am, Coll. Radiol. (2013) 125–143.
[25] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016: pp. 770–778.
[26] H. Wang, J. Feng, Z. Zhang, H. Su, L. Cui, H. He, L. Liu, Breast mass classification via deeply integrating the contextual information from multi-view data, Pattern Recognit. 80 (2018) 42–52.
[27] N.D. Mercaldo, K.F. Lau, X.H. Zhou, Confidence intervals for predictive values with an emphasis to case–control studies, Stat. Med. 26 (2007) 2170–2183.
[28] N. Mao, H. Zhang, Y. Dai, Q. Li, F. Lin, J. Gao, T. Zheng, F. Zhao, H. Xie, C. Xu, H. Ma, Attention-based deep learning for breast lesions classification on contrast enhanced spectral mammography: a multicentre study, Br. J. Cancer. 128 (2023) 793–804, https://doi.org/10.1038/s41416-022-02092-y.
[29] M. Bobowicz, M. Rygusik, J. Buler, R. Buler, M. Ferlin, A. Kwasigroch, E. Szurowska, M. Grochowski, Attention-based deep learning system for classification of breast lesions—multimodal, Weakly Supervised Approach, Cancers. 15 (2023) 2704, https://doi.org/10.3390/cancers15102704.
[30] L. Shen, L.R. Margolies, J.H. Rothstein, E. Fluder, R. McBride, W. Sieh, Deep learning to improve breast cancer detection on screening mammography, Sci. Rep. 9 (2019) 12495, https://doi.org/10.1038/s41598-019-48995-4.
[31] G. Ayana, K. Dese, Y. Dereje, Y. Kebede, H. Barki, D. Amdissa, N. Husen, F. Mulugeta, B. Habtamu, S.-W. Choe, Vision-transformer-based transfer learning for mammogram classification, Diagn. Basel Switz. 13 (2023) 178, https://doi.org/10.3390/diagnostics13020178.
[32] D. Bau B. Zhou A. Khosla A. Oliva A. Torralba Network dissection: quantifying interpretability of deep visual Representations 2017.
[33] F. Hohman, H. Park, C. Robinson, D.H. Polo Chau, Summit: scaling deep learning interpretability by visualizing activation and attribution summarizations, IEEE Trans. vis. Comput. Graph. 26 (2020) 1096–1106, https://doi.org/10.1109/TVCG.2019.2934659.