

Exploring the Limit of Image Resolution for Human Expert Classification of Vascular Ultrasound Images in Giant Cell Arteritis and Healthy Subjects: The GCA-US-AI Project

Claus-Juergen Bauer^{1*}, Stavros Chrysidis², Christian Dejaco^{3,4}, Matthew J Koster⁵, Minna J Kohler⁶, Sara Monti⁷, Wolfgang A Schmidt⁸, Chetan B Mukhtyar⁹, Pantelis Karakostas¹, Marcin Milchert¹⁰, Cristina Ponte^{11,12}, Christina Duftner¹³, Eugenio de Miguel¹⁴, Alojzija Hocevar¹⁵, Annamaria Iagnocco¹⁶, Lene Terslev^{17,18}, Uffe Møller Døhn¹⁸, Berit Dalsgaard Nielsen^{19,20,21}, Aaron Juche²², Luca Seitz²³, Kresten Krarup Keller^{19,20}, Rositsa Karalilova²⁴, Thomas Daikeler²⁵, Sarah Louise Mackie^{26,27}, Karina Torralba²⁸, Kornelis S.M. van der Geest²⁹, Dennis Boumans³⁰, Philipp Bosch³, Alessandro Tomelleri³¹, Markus Aschwanden³², Tanaz A Kermani³³, Andreas Diamantopoulos³⁴, Ulrich Fredberg^{35,36}, Nevsun Inanc³⁷, Simon M Petzinna¹, Shadi Albarqouni^{38,39}, Charlotte Behning⁴⁰, Valentin Sebastian Schäfer¹

1 Clinic of Internal Medicine III, Department of Rheumatology and Clinical Immunology, University Hospital Bonn, Bonn, Germany. 2 Department of Rheumatology, University Hospital of Southern Denmark, Esbjerg, Denmark. 3 Department of Rheumatology and Immunology, Medical University Graz, Graz, Austria. 4 Department of Rheumatology, Hospital of Bruneck (ASAA-SABES), Teaching Hospital of the Paracelsus Medical University, Bruneck, Italy. 5 Division of Rheumatology, Mayo Clinic, Rochester, Minnesota, USA. 6 Division of Rheumatology, Allergy and Immunology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA. 7 IRCCS Istituto Auxologico Italiano, Immunorheumatology Research Laboratory, Milan, Italy 8 Rheumatology, Immanuel Krankenhaus Berlin, Medical Centre for Rheumatology Berlin-Buch, Berlin, Germany. 9 Vasculitis service, Rheumatology department, Norfolk and Norwich University Hospital, Norwich, UK. 10 Department of Rheumatology, Internal Medicine, Geriatrics and Clinical Immunology, Pomeranian Medical University in Szczecin, Szczecin, Poland. 11 Department of Rheumatology, ULS de Santa Maria, Centro Académico de Medicina de Lisboa, Lisbon, Portugal. 12 Faculdade de Medicina, Universidade de Lisboa, Centro Académico de Medicina de Lisboa, Lisbon, Portugal. 13 Department of Internal Medicine, Clinical Division of Internal Medicine II, Medical University Innsbruck, Innsbruck, Austria. 14 Department of Rheumatology, Hospital Universitario La Paz, Madrid, Spain. 15 Department of Rheumatology, University Medical Centre Ljubljana, Ljubljana, Slovenia. 16 Università degli Studi di Torino, Turin, Italy. 17 Department of Clinical Medicine, Copenhagen University, Copenhagen, Denmark. 18 Copenhagen Center for Arthritis Research (COPECARE), Center for Rheumatology and Spine Diseases, Rigshospitalet, Glostrup, Denmark. 19 Department of Clinical Medicine, Aarhus University, Aarhus, Denmark. 20 Department of Rheumatology, Aarhus University Hospital, Aarhus, Denmark. 21 Department of Medicine, The Regional Hospital in Horsens, Horsens, Denmark. 22 Department of Rheumatology, Immanuel Hospital, Berlin, Germany. 23 Department of Rheumatology and Immunology, Inselspital, Bern University Hospital,

University of Bern, Switzerland. 24 Medical University of Plovdiv, Clinic of Rheumatology, University Hospital “Kaspela”, Plovdiv, Bulgaria. 25
Clinic for Rheumatology, University Hospital Basel, Basel, Switzerland. 26 Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of
Leeds, Leeds, UK. 27 Leeds Biomedical Research Centre, Leeds Teaching Hospitals NHS Trust, Leeds, UK. 28 Division of Rheumatology, Department
of Medicine, Loma Linda University School of Medicine, Loma Linda, California, USA. 29 Rheumatology and Clinical Immunology, University of
Groningen, University Medical Center Groningen, Groningen, The Netherlands. 30 Department of Rheumatology and Clinical Immunology, Hospital
Group Twente, Almelo, The Netherlands. 31 Unit of Immunology, Rheumatology, Allergy and Rare Diseases, San Raffaele Scientific Institute, Milan,
Italy. 32 Department of Angiology, University Hospital Basel, Basel, Switzerland. 33 Rheumatology, David Geffen School of Medicine, University of
California Los Angeles, Los Angeles, California, USA. 34 Department of Infectious Diseases, Akershus University Hospital, Lørenskog, Norway. 35
Regional Hospital Northern Jutland, Hjørring, Denmark. 36 Department of Rheumatology, Odense University Hospital, Odense, Denmark. 37 Division
of Rheumatology, Marmara University School of Medicine, Istanbul, Turkey. 38 Department of Diagnostic and Interventional Radiology, University
Hospital Bonn, Bonn Germany. 39 Helmholtz Munich, Helmholtz AI, Neuherberg Germany. 40 Institute of Medical Biometry, Informatics and
Epidemiology, University Hospital of Bonn, Bonn, Germany.

*** Correspondence:**

Corresponding Author

claus-juergen.bauer@ukbonn.de

Keywords: Giant cell arteritis; ultrasound; imaging; image resolution; artificial intelligence; neural
network; image classification

Word count: 3877. Abstract: 249 words. Number of figures: 2+ 2 supplementary. Number of tables:
2.

51 **Key messages:**

52 **WHAT IS ALREADY KNOWN ON THIS TOPIC**

- 53 • Vascular ultrasound is clinically considered the method of choice for the rapid detection of
54 giant cell arteritis; however expertise in ultrasound performance is scarce.
- 55 • Artificial intelligence (AI) could assist in image classification, thereby simplifying the
56 examination process.
- 57 • However, to develop such artificial intelligence, as many AI architectures are constrained by
58 an upper limit of processable image resolution, it is necessary to define the minimum
59 resolution of ultrasound images required to ensure the detection of disease-specific
60 abnormalities.

61 **WHAT THIS STUDY ADDS**

- 62 • 7,750 classification decisions made by 31 international experts were analyzed and found
63 224x224 pixels to be the minimum image resolution required for reliable classification of
64 vascular ultrasound images by human experts.
- 65 • As this result aligns with the identical, critical processing limit for many AI architectures, it
66 allows a wide variety of AI architecture options to be considered for AI development.

67 **HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY**

- 68 • Findings of this study lay the essential foundation for developing an AI system to assist in
69 ultrasound-based detection of giant cell arteritis; an approach holding great potential in
70 allowing rapid diagnosis before ischemic complications occur and improving examination
71 accessibility even in resource-limited regions.

72

73

1 Abstract

Objectives: Prompt diagnosis of giant cell arteritis (GCA) with ultrasound is crucial for preventing severe ocular and other complications, yet expertise in ultrasound performance is scarce. Development of an artificial intelligence (AI)-based assistant facilitating ultrasound image classification and helping to diagnose GCA early promises to close the yet-existing gap. In projection onto the planned AI, this study investigates the minimum image resolution required for human experts to reliably classify ultrasound images of arteries commonly affected by GCA for the presence or absence of GCA.

Methods: Thirty-one international experts in GCA ultrasonography participated in a web-based exercise. They were asked to classify ten ultrasound images per each of five vascular segments as GCA, normal or not able to classify. The following segments were assessed: 1) superficial common temporal artery, 2) its frontal, and 3) parietal branch (all in transverse view), 4) axillary artery in transverse view, and 5) axillary artery in longitudinal view. Identical images were shown at different resolutions, namely 32x32, 64x64, 128x128, 224x224 and 512x512 pixels, thereby resulting in a total of 250 images to be classified by every study participant.

Results: Classification performance improved with increasing resolution up to a threshold, plateauing at 224x224 pixels. At 224x224 pixels, overall classification sensitivity was 0.767 (CI: 0.737–0.796), and specificity was 0.862 (CI: 0.831–0.888).

Conclusions: A resolution of 224x224 pixels ensures reliable human expert classification and aligns with the input requirements of many common AI base architectures. Thus, results of this study guide the projected AI development substantially.

2 Introduction

Giant cell arteritis (GCA) is the most common form of systemic vasculitis in adults and typically affects large and medium size arteries, such as the aorta, temporal and axillary arteries. Signs and symptoms of GCA demand urgent action to verify or exclude the diagnosis in order to prevent severe complications, such as permanent vision loss, which occurs in an average of 20% of all cases^{1,2}, ischemic strokes^{3–5}, and arterial aneurysm and/or dissection⁶. Temporal artery biopsy (TAB) has

historically been the diagnostic gold standard but is increasingly being replaced by ultrasound of the temporal and axillary artery as the first-line diagnostic modality in many parts of the world, a practice supported by the latest European Alliance of Associations for Rheumatology (EULAR) recommendations on imaging for GCA⁷. Ultrasound offers widespread use and rapid availability, patient-friendliness and cost-effectiveness⁸, but also excellent image resolution of less than 0.1mm for superficial arteries with modern ultrasound transducers and the ability to evaluate the entire temporal artery and its branches⁹. This is in contrast to TAB, which may miss GCA detection due to patchy involvement of the artery by vasculitis. The arrival of modern handheld ultrasound devices, characterized by high affordability (<EUR3000), could further expand future access to point-of-care diagnostic imaging, particularly in resource-limited settings¹⁰. While their image quality still has some limitations that depend on the specific application, these constraints are expected to decrease with ongoing technological advancements. As a result, this advancement also has the potential to create new educational opportunities for training both specialists and non-specialists in scanning and interpreting arterial findings in the context of GCA.

GCA ultrasound has proven its excellent reliability^{11–16} among experts using appropriate ultrasound equipment, defined as a linear probe ≥ 10 MHz for axillary artery assessment and a linear probe ≥ 18 MHz for cranial artery evaluation¹⁷. However, the examination, and especially the interpretation of findings, requires experience and expertise^{14,17–19} that are rare to find and mostly limited to specialized centres. While there are initiatives to train rheumatologists all around the globe, these efforts will presumably be insufficient to achieve the desirable goal of a global network of fast-track clinics with sufficient expertise in GCA ultrasound. With modern technological advances, a suggested solution to close this gap is the development of an artificial intelligence (AI)-based assistant (e.g. publicly accessible online) that can help classify submitted ultrasound images as abnormal or not, thereby assisting physicians globally in the early detection of GCA.

AI has been proclaimed to lead the fourth industrial revolution and is seen as having the potential of heavily impact our present and future²⁰. Likewise, the introduction of AI in the healthcare sector has accelerated quickly over the last few years²¹. Previous attempts to apply AI in the classification of medical image data have proven feasibility and shown remarkable results when trained to label chest radiographs²², classify skin cancer²³ or identify breast cancer in mammograms²⁴. Recently, the first applications of AI in the field of imaging in rheumatology have been made: Andersen et al. developed a neural network for automatic scoring of arthritis disease activity on joint ultrasound images²⁵, and Bressemer et al. trained neural networks for the accurate detection of specific inflammatory joint

abnormalities (“sacroiliitis”) in pelvic x-rays²⁶. However, no attempt has yet been made to utilize AI in GCA for classifying B-mode ultrasound images—standard two-dimensional grayscale ultrasound images without color Doppler—as normal or abnormal.

The technological foundation of AI is marked by deep neural networks²⁷. A variety of different deep neural network architectures exist, each with distinct characteristics, including limitations in image processing size, resolution, result accuracy, and resource efficiency²⁷. In pursuing the outlined development of an artificial intelligence-based system (see **Supplementary Material 1: Conception of the envisioned AI**), it is crucial to consider the technological constraints of diverse neural network architectures. Neural network architectures that are not able to handle images presented in a sufficiently high resolution necessary for the adequate evaluation of disease-characteristic features should be excluded in advance. In this context, the image resolution of 224x224 pixels represents a critical threshold, as it is the maximum limit for input in many common neural network architectures^{26,28}. On the other hand, neural network architectures allowing for higher resolution image input are available, but the processing of higher image resolutions increases computing power demand. Limitations in the available computing power require specifications limiting the maximum image resolution intended for processing. This ensures the selection of a suitable deep neural network architecture that balances resource efficiency with the ability to handle task-appropriate image sizes. In order to provide scientific evidence and determine the requirements a neural network must fulfill, this project aims to define the minimum resolution required for human experts to reliably classify ultrasound images of arteries commonly affected by GCA for the presence or absence of the disease.

3 Materials and Methods

3.1 Study design

We designed a web-based exercise in which physicians experienced in GCA ultrasonography were asked to classify 250 ultrasound images for the presence or absence of GCA-typical morphology. Forty-one physicians from 15 countries were invited by email to participate. To qualify as experts, all participants were board-certified rheumatologists with extensive experience in large-vessel vasculitis ultrasonography, defined as a minimum of four years interpreting GCA ultrasound images and prior

assessment of at least 100 patients with suspected GCA throughout their careers. Additionally, they were members of the Large Vessel Vasculitis subgroup within the Outcome Measures in Rheumatology (OMERACT) Ultrasound subgroup. The study followed the principles of the Declaration of Helsinki and the Guidelines of Good Clinical Practice and received ethical approval by the local ethics committee (IRB No. #101/22). Patients and public were not actively involved in design and conduct of this study.

3.2 Vascular ultrasound image source and selection

CJB selected 50 ultrasound images from an image collection provided by five international experts in GCA ultrasonography (CM, CP, MM, VSS and WAS). The selection included 10 images of each of the following vessels: 1) Superficial common temporal arteries, 2) its frontal branch, 3) its parietal branch (all in transverse view), 4) axillary arteries in transverse view and 5) axillary arteries in longitudinal view. All ultrasound images were obtained in B-mode. The sample included six images from patients with acute GCA and four images from healthy individuals for each vessel, except for the axillary artery in longitudinal view, which had five images from patients with acute GCA and five images from healthy individuals. The images were obtained from a total of 50 patients, of which 29 had GCA and 21 had no evidence of GCA.

Each image was derived from an individual patient or healthy subject. Images were obtained from different ultrasound machine brands (Esaote, GE and Philips) using 3-12 (Philips Epiq), 4-12 (GE LOGIQ e), 6-15 (GE Logiq S8), 5-18 (Philips Epiq), 6-18 (Esaote MyLab 8), 8-18 (GE Logiq E9) and 6-24 MHz (GE Logiq E10) linear transducers for the investigation of axillary arteries and 5-18 (Philips Epiq), 8-18 (GE Logiq S8, GE Logiq E9) and 6-24 MHz (GE Logiq E10) hockey stick probes for the investigation of cranial arteries. Image acquisition at each site was performed according to a previously published scanning protocol¹⁴. Intima media thickness (IMT) measurements were available for all images. All GCA patients had a diagnosis confirmed by a board-certified rheumatologist and met either the extension of the 2016 American College of Rheumatology classification criteria for GCA²⁹ or the 2022 American College of Rheumatology/EULAR classification criteria for giant cell arteritis³⁰ (depending on the date of diagnosis). Diagnosis was confirmed either by TAB or imaging, including ultrasound. None of the patients had a change in clinical diagnosis from GCA to a non-GCA diagnosis during the follow- up period.

193

194 **3.3 Vascular ultrasound image processing**

195 All ultrasound images were individually loaded into an image editing software (GNU Image
196 Manipulation Program version 2.10.8) to apply the necessary image preparation for this study. IMT
197 measurements were removed if present. From each ultrasound image, a square 512x512-pixel image
198 section focusing on the appropriate artery was selected and cut out. This section was subsequently
199 downsized (without interpolation) to all study-specific resolution variants, namely 32x32, 64x64,
200 128x128 and 224x224 pixels. **Figure 1** reflects the visual perception of all five image resolution
201 variants applied in the study. All four downsized variants, in addition to the initial 512x512 pixel
202 variant, resulted in a total of 250 vascular ultrasound images representing 50 distinct vessel images,
203 each at five different resolutions.

204

205 **3.4 Conduct of the web-based classification exercise**

206 The web-based image classification exercise was performed via Research Electronic Data Capture
207 (REDCap, provided by Vanderbilt University³¹) version 13.2.5. It included all vascular ultrasound
208 images without any information about the corresponding IMT measurements, final ultrasound result
209 or clinical patient details. Intentionally, participants of the exercise were not told that they were
210 evaluating identical images at different resolutions. To prevent image recall, images were displayed
211 sequentially by artery and section, starting with the lowest resolution, with images presented in random
212 order at each resolution level. During the classification exercise, each image was shown individually
213 (one per page), with each vascular ultrasound image representing a separate patient case (see
214 Supplementary **Figure 1**). Participants were given three response options to the question “I classify
215 this ultrasound image as belonging to a ...”: 1) “GCA patient”, 2) “Healthy individual” and 3)
216 “Meaningful classification is not possible”. Whenever the latter option was selected, a follow-up dialog
217 box appeared, prompting the participant to choose either “GCA patient” or “Healthy individual” in
218 response to the question, “If only one of the above-mentioned answers was accepted, I would pick...”.
219 Personal data regarding professional experience was also collected from all participants.

220

221 **3.5 Statistical analysis**

Statistical analyses were conducted using SPSS ® Statistics (Version 26.0.0.0; IBM Corp., Armonk, USA) and R (Version 4.3.0; R Foundation for Statistical Computing, Vienna, Austria)³². Descriptive statistics and explorative data analysis were used. For metric parameters mean, standard deviation, median (in the case of the monthly average image classification figures, which contained individual outliers) and ranges were calculated. Categorical data were summarized by absolute and relative frequencies. Further analysis to determine to what extent a change in image resolution is linked with a significant effect on correct image classification was performed using logistic mixed-effects models. An increased chance of correct classification, indicated by an estimated odds ratios (OR) larger than 1, is interpreted as improvement of classification performance. An estimated OR smaller than one is interpreted as deterioration in image classification performance. To account for repeated measurements per rater and image, these were included as random intercept terms. Further factors impacting the classification decision, such as image resolution, working experience and whether a study participant had provided study images, were investigated and therefore included as fixed effects. The image resolution and arteries were included as a categorical independent variables. When the analyses were adjusted for the number of years experts had already been reading GCA ultrasound images or the number of patients that experts estimated to have seen, these were included as metric independent variables. Sensitivity and specificity were calculated on the basis of correct and incorrect classification numbers. In addition to that, 95% confidence intervals (CI) were calculated. Due to a technical error in REDCap, detected only after study completion, participants could bypass the ‘forced classification’ within the follow-up question after the initial response had indicated an unclassifiable image. This affected all participants equally. Missing classifications from this issue were excluded from the analysis. The threshold for statistical significance was set at $p=0.05$.

4 Results

4.1 Study participant characteristics and expertise

Out of 41 international experts in GCA ultrasonography invited to participate in the web-based exercise, 31 responded, resulting in a response rate of 75.6%. The cohort comprised five board-certified rheumatologists from Denmark, four from the United States of America, four from Germany, three from Italy, three from Austria, two from the United Kingdom, two from The Netherlands and one each

from Bulgaria, Norway, Poland, Portugal, Slovenia, Spain, Switzerland, and Turkey. Study participants had an average of 21.2 years of medical experience, with a range from 9 to 40 years. They had been interpreting GCA ultrasound images for an average of 10.9 years, with experience ranging from 4 to 30 years. Participants reported to evaluate a median of 16.0 (mean: 70.1, range 2-500) images from GCA patients per month.

4.2 Image classification results

Overall, 7,750 classification decisions were obtained (31 participants conducting a total of 310 decisions per artery, view and resolution combination).

Overall, the frequency participants selected the response option “meaningful classification is not possible” decreased with increasing image resolution (constantly with every sequential stage of raised image resolution) from 90.5% (at the image resolution stage of 32x32 pixels) to 25.9% (at 512x512 pixels). **Figure 2** illustrates how frequently the response option was selected during vascular ultrasound image classification depending on the respective artery (and view) and image resolution. It was selected least frequently for images of the axillary artery in longitudinal view (19.0%), the axillary artery in transverse view (22.3%), and the temporal artery in transverse view (26.1%) at the 512x512 pixel resolution.

When experts initially selected “meaningful classification is not possible”, they were subsequently asked to make a forced decision on whether the ultrasound image belonged to a GCA patient or a healthy individual (referred to as 'forced classification'). **Table 1** presents detailed results of 'confident classification' provided in the first instance, as well as the 'forced classification' outcomes for all arteries, views, and image resolutions. Generally, increasing image resolution from 32x32 to 224x224 pixels led to a higher number and percentage of correct classifications across all arteries and views. This improvement was consistent whether considering only correct confident classifications (32x32 pixels: 4.8%; 64x64 pixels: 13.8%; 128x128 pixels: 37.7%; 224x224 pixels: 57.6%). or both confident and forced classifications (32x32 pixels: 37.8%; 64x64 pixels: 54.8%; 128x128 pixels: 68.6%; 224x224 pixels: 74.4%). No relevant increment was observed in classification results once image resolution increased further from 224x224 pixels to 512x512 pixels (overall correct confident classification at 512x512 pixels: 61.1%; overall correct confident and forced classification at 512x512 pixels: 75.3%).

Detailed view showed the highest percentage of correct classifications was achieved in ultrasound images of the axillary artery in longitudinal view (total percentage of correct classifications at an image resolution of 512x512 pixels: 83.9%). Images of the axillary artery in transverse view yielded the lowest percentage of correct classifications at any image resolution stage compared to other combinations of arteries and views.

Further analysis of the complete dataset across all arteries and views revealed that any increase in image resolution significantly improved classification performance when compared to the reference category of 32x32 pixel images. Specifically, image resolutions of 64x64 pixels (OR 2.19, CI 1.84–2.61, $p < 0.001$), 128x128 pixels (OR 3.49, CI 2.91–4.19, $p < 0.001$), 224x224 pixels (OR 4.11, CI 3.42–4.94, $p < 0.001$), and 512x512 pixels (OR 3.90, CI 3.25–4.69, $p < 0.001$) were all associated with significant improvements in image classification performance. In comparison to 128x128 pixel images as the reference category, increasing the image resolution further improved classification performance numerically, although the improvement did not reach statistical significance (224x224 pixels: OR 1.18, CI 0.97–1.43, $p = 0.094$; 512x512 pixels: OR 1.12, CI 0.93–1.35, $p = 0.248$). A sub-analysis selectively considering images of the axillary artery in longitudinal view or the temporal artery in transverse view did not reveal any relevant differences in the overall trend as compared to the main analysis.

Heatmap plots illustrating the frequency each individual image included in this web-based exercise was classified correctly at a certain image resolution level are enclosed as **Supplementary Figure 2**.

4.3 Sensitivity and specificity

The sensitivity and specificity of classifying vascular ultrasound images for the presence or absence of GCA-indicative abnormalities based on single static B-mode images was assessed for any vessel/view and image resolution. Classification of 512x512 pixel longitudinal view images of the axillary artery achieved the highest specificity (specificity: 0.948, sensitivity: 0.734) while the lowest specificity was observed in 64x64 pixel transverse view images of the axillary artery (specificity: 0.596). Classification of 512x512 pixel transverse view images of the temporal artery's frontal branch achieved the highest sensitivity (sensitivity: 0.851, specificity: 0.793) while the lowest sensitivity was observed in 32x32 pixel transverse view images of the parietal branch of the temporal artery (sensitivity: 0.153). In general, sensitivity and specificity metrics constantly improved with rising image resolution between 32x32 and 128x128 pixels. In further increasing the image resolution beyond 128x128 pixels,

sensitivity and specificity metrics continued to rise or remained constant in dependence of the considered artery and view. Detailed data is listed in **Table 2**.

4.4 Factors impacting the classification decision

Logistic mixed-effects model, that was performed in order to detect influential factors in classification decision-making and potential confounders (study participants' contribution to study image provision, image resolution, study participants' level of experience in GCA image interpretation as quantified by temporal duration and quantity of scans interpreted throughout the career), identified image resolution as the only significant determining factor. A lower image resolution had a higher chance that the image was marked as "meaningful classification is not possible". Compared to a resolution of 512x512 pixels, the chance for receiving the label "Meaningful classification is not possible" (1: yes, 0: no) was increased by the factor 1.43 ($p < 0.001$) for images of size 224x224 pixels, by the factor 5.05 ($p < 0.001$) for images of size 128x128 pixels, by the factor 40.97 ($p < 0.001$) for images of size 64x64 pixels, and by the factor 198.84 ($p < 0.001$) for an image size of 32x32 pixels. Likewise, compared to a resolution of 512x512 pixels, lower image resolutions led to significantly lower rates of correct image classification (e.g. for an image size of 32x32 pixels: OR = 0.25, $p < 0.001$; for an image size of 64x64 pixels: OR = 0.56, $p < 0.001$). The number of years experts had already been reading GCA ultrasound images had no significant effect on the selection of the response category "Meaningful classification is not possible" (OR= 1.02, $p = 0.725$). Neither did the number of patients that experts estimated to have seen and evaluated for suspected GCA throughout their career (OR= 1.00, $p = 0.422$). Presented statistical data was derived from a model adjusting for differences across artery locations and views.

5 Discussion

In this first and international study exploring the minimum image resolution mandatory for reliable human expert classification of vascular ultrasound images in GCA and healthy subjects, we observed ascending classification performance with increasing image resolution up to an image resolution of 224x224 pixels where decision confidence and image classification performance plateaued without further significant gains despite a continual increase in image resolution.

In fact, an in-depth analysis of the classification exercise performed in this study by human experts revealed that compared to images with 128x128 pixels - when set as the reference category - higher image resolutions still showed further numerical improvement in classification performance. However, classification performance improvement with increasing image resolution largely plateaued at 224x224 pixels, in line with sensitivity and specificity metrics that also accentuated a ‘sweet spot’ at 224x224 pixels. In a combined conservative consideration of these results, an image resolution of 224x224 pixels crystallized as the optimum between preserved information content and data size efficiency, thereby defining the minimum requirement in image resolution a neural network architecture in this scenario of application will be supposed to handle. These findings resonate with the applied image resolution in other application scenarios of AI-driven medical imaging classification^{33,34} in general, but more specifically vessel ultrasound image classification within the scope of other disorders than GCA^{35,36}. Particularly noteworthy is a study conducted by Li et al³⁹, which similarly focused on the detection of disease-specific vascular wall changes - in this case the automatic detection of atherosclerotic plaque and calcification from intravascular ultrasound images. Despite contextual differences, parallels to our study’s conclusions can be drawn from the fact that Li et. al also successfully trained their neural network with images in a resolution of 224×224 pixels. Our study provides evidence that adequate GCA ultrasound image classification by human experts does not necessarily require the highest resolution level of 512x512 pixels but can be appropriately achieved at a resolution level of 224x224 pixels. This finding broadens the range of neural network architectures suitable for artificial intelligence development, including the widely chosen and resource-efficient network architectures such as ResNet²⁶ or DenseNet²⁸, which are limited to a maximum default image input size of 224x224 pixels.

Moreover, the reported results on human expert image classification performance serve as an essential benchmark for the classification performance of the projected neural network, which aims to match or exceed human expert classification performance, as achieved in other medical fields before³⁴. The data presented in this study is representative, particularly considering that experts were only provided with one single static B-mode image per patient without any additional IMT measurement, halo, compression sign, insight into other anatomical segments or clinical information. While participants in this study classified each image individually, it is important to note that there are typically eight or more images available per patient, which may lead to higher sensitivity and lower specificity on a patient level when combined in clinical practice. Also, sensitivity and specificity results, combined with heatmap visualization (Supplementary Figure 2) and data from both confident and forced

classifications, indicate that participants tended to classify an image as belonging to a GCA patient only when they were relatively confident. Among all cases in which the initial response option had been “meaningful classification is not possible”, subsequent forced classifications found more than twice as many of these images to be attributed to GCA patients versus healthy individuals. Thus, images ultimately classified to belong to GCA patients were clearly overrepresented within the category of “meaningful classification is not possible”-initial responses. This superordinate dynamic likely contributed to missed GCA classifications, explaining the persistent rate of misclassifications even at higher image resolutions. At the same time, this caused assessment specificity to remain consistently high across all resolution stages. Despite all these factors, the image classification performance yielded remarkable sensitivity and specificity results. Our reported sensitivity of 0.833 and specificity of 0.915 for 224x224 pixel longitudinal view images of the axillary artery, as well as 0.734 and 0.855, respectively, for 224x224 pixel transverse view images of the superficial common temporal artery, align with data from a systematic review and meta-analysis investigating the diagnostic performance of temporal artery ultrasound (considering full resolution images) for the diagnosis of GCA³⁷. This meta-analysis, based on eleven studies with available data, reported a sensitivity of 0.78 and a specificity of 0.79 for detecting any sono-morphologic abnormalities (hypoechoic halo, stenosis and/or occlusion) in GCA ultrasound. Likewise, this study’s results are in line with the results of the OMERACT Patient-based Reliability Exercise¹⁴ that yielded an interreader reliability of around 0.8 (Light κ 0.76–0.86) for the overall diagnosis of GCA.

Limitations included the natural restrictions in expert availability to participate in this exercise. Indeed, five experts participated in this study that had previously granted access to their image libraries out of which a fraction of images was randomly selected by an independent individual (CJB), not involved as a study participant. In weighing whether these study participants may have contributed to a bias in the results, we performed sensitivity analyses for all reported metrics, which did not lead to a significant change in results after the exclusion of these participants. Image selection may have been subject to a selection bias and holds the potential to have impaired representativeness. Furthermore, we decided neither to actively include nor exclude ultrasound images containing aspects of atherosclerosis, where an OMERACT ultrasound definition is available and validated¹⁵. Instead, our dataset represents a random sampling from real-world clinical cases, **not stratified based on comorbidities**. While, in fact, fibromuscular dysplasia or many other vasculopathies are rarely encountered in temporal and axillary

arteries and also atherosclerosis is uncommon in temporal arteries, atherosclerotic changes can occur in the axillary arteries, though are typically distinguishable by ultrasound based on their characteristic morphology¹⁵. Nonetheless, future research should explore whether higher image resolution is needed to differentiate GCA from these specific mimics, particularly in borderline cases where classification might be more challenging. Intentionally and with the operating principle of the projected AI-based GCA ultrasound image classification assistant in mind (see **Supplementary Material 1: Conception of the envisioned AI**), this study made use of one single static B-mode image per patient case without any additional IMT measurement, halo, compression sign or clinical information. This represents a striking difference in the way human experts are used to approach a classification decision in clinical practice when performing GCA ultrasound in a dynamic manner with appraisal of multiple anatomical segments. We regard this as the primary reason for the overall high rate of initially “non-classifiable” images. An intra-rater exercise could also have been conducted but was felt to be beyond the scope of this project. A technical error in REDCap allowed the omission of forced classifications, resulting in missing values within the dataset. While potential bias cannot be entirely excluded, removing these missing classifications was deemed the most appropriate strategy to minimize bias. For transparency, **Table 1** presents a detailed breakdown of classification results across all arteries, views, and image resolutions for both confident and forced classifications.

In conclusion, this international online study evaluated the minimum image resolution mandatory for reliable human expert classification of vascular ultrasound images in GCA and healthy subjects and found that 224x224 pixel images provide the optimum between preserved information content and data size efficiency. This opens up a broad variety of applicable deep neural network architectures for the development of an artificial intelligence-based assistant. Building on this, a 3,800 ultrasound image library is already provided for deep neural network training that will be followed up by clinical validation to proof reliability. Consecutively, a publicly accessible website hosting the AI-based assistant that facilitates classification of uploaded ultrasound images and helps to diagnose GCA early is envisioned. This approach holds great potential for the rapid and accurate diagnosis of GCA including in resource-disadvantaged regions; a factor that could critically impact individuals by allowing diagnosis before ischemic complications occur.

434

435

436 **6 Conflict of Interest**

437 The authors declare that the research was conducted in the absence of any commercial or financial
438 relationships that could be construed as a potential conflict of interest.

439

440 **7 Author Contributions [CRediT author statement]**

441 **Claus-Juergen Bauer:** Conceptualization, methodology, validation, formal analysis, investigation,
442 resources, data curation, writing (original draft), writing (review and editing), visualization,
443 supervision, project administration

444 **Stavros Chrysidis:** Methodology, software, validation, investigation, resources, data curation,
445 writing (review and editing), funding acquisition

446 **Wolfgang A Schmidt, Marcin Milchert, Cristina Ponte, Chetan B Mukhtyar:** Investigation,
447 resources, writing (review and editing)

448 **Christian Dejaco, Matthew J Koster, Minna J Kohler, Sara Monti, Wolfgang A Schmidt,**
449 **Pantelis Karakostas, Marcin Milchert, Cristina Dias Botelho da Ponte, Chetan B Mukhtyar,**
450 **Christina Duftner, Eugenio de Miguel, Alojzija Hocevar, Annamaria Iagnocco, Lene Terslev,**
451 **Uffe Møller Døhn, Berit Dalsgaard Nielsen, Aaron Juche, Luca Seitz, Kresten Krarup Keller,**
452 **Rositsa Karalilova, Thomas Daikeler, Sarah Louise Mackie, Karina Torralba, Kornelis S.M.**
453 **van der Geest, Dennis Boumans, Philipp Bosch, Alessandro Tomelleri, Markus Aschwanden,**
454 **Tanaz A Kermani, Andreas Diamantopoulos, Ulrich Fredberg, Nevsun Inanc, Simon M**
455 **Petzinna, Shadi Albarqouni:** Investigation, writing (review and editing)

456 **Charlotte Behning:** Formal analysis, resources, writing (original draft), writing (review and editing),
457 visualization

458 **Valentin Sebastian Schäfer:** Methodology, investigation, resources, writing (original draft), writing
459 (review and editing), supervision, internal funding acquisition

460

461 **8 Funding**

462 This study received no specific funding.

463

464 **9 Acknowledgments**

465 SLM reports: SLM is supported in part by the NIHR Leeds Biomedical Research Centre
466 (NIHR203331). The views expressed in this article are those of the authors and not necessarily those
467 of the NIHR, the NIHR Leeds Biomedical Research Centre, the National Health Service or the UK
468 Department of Health and Social Care.

469

470 **10 Data Availability Statement**

471 Any dataset parts generated and analyzed for this study that are not already included in this article,
472 are available upon reasonable request. Further inquiries can be directed to the corresponding author.

473

474 **11 References**

- 475 1. Aiello PD, Trautmann JC, McPhee TJ, et al. Visual prognosis in giant cell arteritis.
476 *Ophthalmology*. 1993;100:550–5.
- 477 2. Salvarani C, Cimino L, Macchioni P, et al. Risk factors for visual loss in an Italian population-
478 based cohort of patients with giant cell arteritis. *Arthritis and rheumatism*. 2005;53:293–7.
- 479 3. Pariente A, Guédon A, Alamowitch S, et al. Ischemic stroke in giant-cell arteritis: French
480 retrospective study. *Journal of autoimmunity*. 2019;99:48–51.
- 481 4. Caselli RJ, Hunder GG. Neurologic complications of giant cell (temporal) arteritis. *Seminars in*
482 *neurology*. 1994;14:349–53.
- 483 5. Gonzalez-Gay MA, Vazquez-Rodriguez TR, Gomez-Acebo I, et al. Strokes at time of disease
484 diagnosis in a series of 287 patients with biopsy-proven giant cell arteritis. *Medicine*.
485 2009;88:227–235.

- 486 6. Gonzalez-Gay MA, Garcia-Porrúa C, Piñeiro A, et al. Aortic aneurysm and dissection in
487 patients with biopsy-proven giant cell arteritis from northwestern Spain: a population-based
488 study. *Medicine (Baltimore)*. 2004;83:335–341.
- 489 7. Hellmich B, Agueda A, Monti S, et al. 2018 Update of the EULAR recommendations for the
490 management of large vessel vasculitis. 2020;79:19–30.
- 491 8. Luqmani R, Lee E, Singh S, et al. The Role of Ultrasound Compared to Biopsy of Temporal
492 Arteries in the Diagnosis and Treatment of Giant Cell Arteritis (TABUL): a diagnostic accuracy
493 and cost-effectiveness study. *Health technology assessment (Winchester, England)*. 2016;20:1–
494 238.
- 495 9. Schmidt WA. Ultrasound in the diagnosis and management of giant cell arteritis. *Rheumatology*.
496 2018;57:ii22–ii31.
- 497 10. Karakostas P, Dejaco C, Behning C, et al. Point-of-care ultrasound enables diagnosis of giant
498 cell arteritis with a modern innovative handheld probe. *Rheumatology*. 2021;60:4434–4436.
- 499 11. Schmidt WA, Kraft HE, Vorpahl K, et al. Color duplex ultrasonography in the diagnosis of
500 temporal arteritis. *The New England journal of medicine*. 1997;337:1336–42.
- 501 12. Aschwanden M, Imfeld S, Staub D, et al. The ultrasound compression sign to diagnose temporal
502 giant cell arteritis shows an excellent interobserver agreement. *Clinical and experimental*
503 *rheumatology*. 2015;33:S-113-5.
- 504 13. De Miguel E, Castillo C, Rodríguez A, et al. Learning and reliability of colour Doppler
505 ultrasound in giant cell arteritis. *Clinical and experimental rheumatology*. 2009;27:S53-8.
- 506 14. Schäfer VS, Chrysidis S, Dejaco C, et al. Assessing Vasculitis in Giant Cell Arteritis by
507 Ultrasound: Results of OMERACT Patient-based Reliability Exercises. *J Rheumatol*.
508 2018;45:1289–1295.
- 509 15. Chrysidis S, Duftner C, Dejaco C, et al. Definitions and reliability assessment of elementary
510 ultrasound lesions in giant cell arteritis: a study from the OMERACT Large Vessel Vasculitis
511 Ultrasound Working Group. 2018;4:e000598.
- 512 16. Chrysidis S, Terslev L, Christensen R, et al. Vascular ultrasound for the diagnosis of giant cell
513 arteritis: a reliability and agreement study based on a standardised training programme. *RMD*
514 *open*. 2020;6:e001337.
- 515 17. Coath FL, Bukhari M, Ducker G, et al. Quality standards for the care of people with giant cell
516 arteritis in secondary care. *Rheumatology*. 2023;62:3075–3083.
- 517 18. Terslev L, Hammer HB, Torp-Pedersen S, et al. EFSUMB minimum training requirements for
518 rheumatologists performing musculoskeletal ultrasound. *Ultraschall in der Medizin (Stuttgart,*
519 *Germany : 1980)*. 2013;34:475–7.
- 520 19. Muratore F, Pipitone N, Salvarani C, et al. Imaging of vasculitis: State of the art. *Best practice*
521 *& research Clinical rheumatology*. 2016;30:688–706.

- 522 20. Chakraborty U, Banerjee A, Saha JK, et al. Artificial intelligence and the fourth industrial
523 revolution. CRC Press; 2022.
- 524 21. Amisha, Malik P, Pathania M, et al. Overview of artificial intelligence in medicine. *Journal of*
525 *family medicine and primary care*. 2019;8:2328–2331.
- 526 22. Irvin J, Rajpurkar P, Ko M, et al. Chexpert: A large chest radiograph dataset with uncertainty
527 labels and expert comparison. 2019:590–597.
- 528 23. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with
529 deep neural networks. *Nature*. 2017;542:115–118.
- 530 24. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast
531 cancer screening. *Nature*. 2020;577:89–94.
- 532 25. Andersen JKH, Pedersen JS, Laursen MS, et al. Neural networks for automatic scoring of
533 arthritis disease activity on ultrasound images. *RMD open*. 2019;5:e000891.
- 534 26. Bressem KK, Vahldiek JL, Adams L, et al. Deep learning for detection of radiographic
535 sacroiliitis: achieving expert-level performance. *Arthritis Research & Therapy*. 2021;23:106.
- 536 27. Alom MZ, Taha TM, Yakopcic C, et al. The history began from alexnet: A comprehensive
537 survey on deep learning approaches. *arXiv preprint arXiv:180301164*.
- 538 28. Riasatian A, Babaie M, Maleki D, et al. Fine-tuning and training of densenet for histopathology
539 image representation using tcga diagnostic slides. *Medical image analysis*. 2021;70:102032.
- 540 29. Salehi-Abari I. 2016 ACR revised criteria for early diagnosis of giant cell (temporal) arteritis.
541 *Autoimmune Dis Ther Approaches*. 2016;3:1–4.
- 542 30. Ponte C, Grayson PC, Robson JC, et al. 2022 American College of Rheumatology/EULAR
543 Classification Criteria for Giant Cell Arteritis. *Arthritis & Rheumatology*. 2022;74:1881–1889.
- 544 31. Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap)--a metadata-
545 driven methodology and workflow process for providing translational research informatics
546 support. *J Biomed Inform*. 2009;42:377–381.
- 547 32. R Core Team. R: A language and environment for statistical computing. R Foundation for
548 Statistical Computing, Vienna, Austria. <http://www.R-project.org/>. Available from:
549 <http://www.R-project.org/>. 2023. Accessed April 17, 2023.
- 550 33. Sabottke CF, Spieler BM. The effect of image resolution on deep learning in radiography.
551 *Radiology: Artificial Intelligence*. 2020;2:e190015.
- 552 34. Zunair H, Hamza AB. Melanoma detection using adversarial training and deep transfer learning.
553 *Physics in Medicine & Biology*. 2020;65:135005.
- 554 35. Xie M, Li Y, Xue Y, et al. Vessel lumen segmentation in internal carotid artery ultrasounds with
555 deep convolutional neural networks. *IEEE*; 2019:2393–2398.

36. Li Y-C, Shen T-Y, Chen C-C, et al. Automatic detection of atherosclerotic plaque and calcification from intravascular ultrasound images by using deep convolutional neural networks. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*. 2021;68:1762–1772.
37. Rinagel M, Chatelus E, Jousse-Joulin S, et al. Diagnostic performance of temporal artery ultrasound for the diagnosis of giant cell arteritis: a systematic review and meta-analysis of the literature. *Autoimmunity reviews*. 2019;18:56–61.

12 Tables

		Confident classification			Forced classification						
Location	Pixel	Correct	Incorrect	Not possible				% confident correct	% total correct*	% confident incorrect	% total incorrect**
				↳	Correct	Incorrect	Not answered				
Axillary artery - Longitudinal view	32x32	18	12	280	92	99	89	5.81	35.48	3.87	35.81
	64x64	38	13	259	146	51	62	12.26	59.35	4.19	20.65
	128x128	124	21	165	112	26	27	40.00	76.13	6.77	15.16
	224x224	201	26	83	49	27	7	64.84	80.65	8.39	17.10
	512x512	227	24	59	33	25	1	73.23	83.87	7.74	15.81
Axillary artery - Transverse view	32x32	12	16	282	98	106	78	3.87	35.48	5.16	39.35
	64x64	49	24	237	97	78	62	15.81	47.10	7.74	32.90
	128x128	113	42	155	70	42	43	36.45	59.03	13.55	27.10
	224x224	172	47	91	35	28	28	55.48	66.77	15.16	24.19
	512x512	176	65	69	29	19	21	56.77	66.13	20.97	27.10
Superficial common temporal artery - Transverse view	32x32	19	18	273	105	85	83	6.13	40.00	5.81	33.23
	64x64	42	27	241	117	64	60	13.55	51.29	8.71	29.35
	128x128	115	28	167	87	41	39	37.10	65.16	9.03	22.26
	224x224	166	43	101	52	18	31	53.55	70.32	13.87	19.68
	512x512	180	49	81	43	16	22	58.06	71.94	15.81	20.97
Frontal branch of the temporal artery - Transverse view	32x32	14	15	281	112	92	77	4.52	40.65	4.84	34.52
	64x64	38	17	255	142	39	74	12.26	58.06	5.48	18.06
	128x128	99	28	183	128	20	35	31.94	73.23	9.03	15.48
	224x224	159	28	123	70	23	30	51.29	73.87	9.03	16.45
	512x512	168	32	110	68	17	25	54.19	76.13	10.32	15.81

Parietal branch of the temporal artery - Transverse view	32x32	11	12	287	105	112	70	3.55	37.42	3.87	40.00
	64x64	47	18	245	134	56	55	15.16	58.39	5.81	23.87
	128x128	133	29	148	82	24	42	42.90	69.35	9.35	17.10
	224x224	194	28	88	55	8	25	62.58	80.32	9.03	11.61
	512x512	196	32	82	47	11	24	63.23	78.39	10.32	13.87

Table 1. Image classification results across all arteries, views and image resolutions [in pixels].

During the image classification exercise, experts first chose one of three options to answer the question, “*I classify this ultrasound image as belonging to a ...*”: (1) “*GCA patient*”, (2) “*Healthy individual*”, or (3) “*Meaningful classification is not possible*”. This *initial decision is termed “confident classification”* (results in the first three columns). If option (3) was selected, a follow-up dialog prompted a forced choice between “*GCA patient*” and “*Healthy individual*” (termed “forced classification”, results in the four right columns). Percentage figures can be found in the four columns on the right. In total, 310 classifications were submitted for each artery, view, and image resolution (10 images classified by 31 experts).

* % total correct= number of correct confident classifications plus number of correct forced classifications, divided by total number of classifications (n=310)

** % total incorrect= number of incorrect confident classifications plus number of incorrect forced classifications, divided by total number of classifications (n=310)

Artery		Image Resolution [Pixel]				
		32x32	64x64	128x128	224x224	512x512
Overall	Sensitivity	0.299	0.585	0.755	0.767	0.761
	Specificity	0.814	0.832	0.814	0.862	0.856
Axillary artery - longitudinal view	Sensitivity	0.328	0.549	0.727	0.733	0.734
	Specificity	0.719	0.929	0.943	0.915	0.948
Axillary artery - transverse view	Sensitivity	0.350	0.584	0.707	0.702	0.678
	Specificity	0.663	0.596	0.65	0.784	0.757

Superficial common temporal artery - transverse view	Sensitivity	0.346	0.549	0.759	0.734	0.709
	Specificity	0.846	0.773	0.724	0.855	0.871
Frontal branch* - transverse view	Sensitivity	0.329	0.690	0.837	0.834	0.851
	Specificity	0.860	0.872	0.807	0.793	0.793
Parietal branch* - transverse view	Sensitivity	0.153	0.555	0.738	0.829	0.830
	Specificity	0.979	0.950	0.898	0.939	0.878

Table 2. Sensitivity and specificity of vascular ultrasound image classification by human experts at various image resolutions, arteries and views.

*Branch of the temporal artery

13 Figure legends

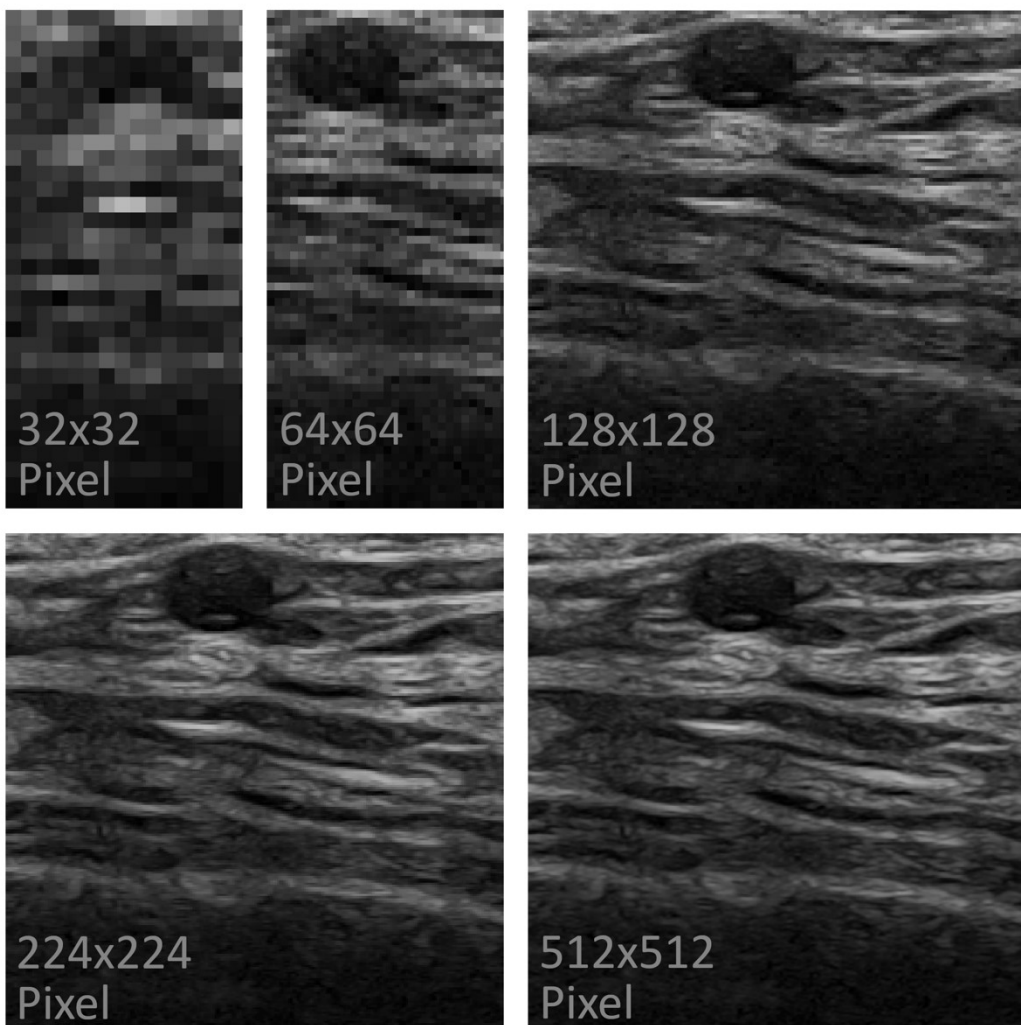


Figure 1. Comparison between the original square 512x512 pixel cut-out from a native B-mode ultrasound image of the superficial common temporal artery in transversal view and its downsized study-specific resolution variants, namely 32x32, 64x64, 128x128 and 224x224 pixels.

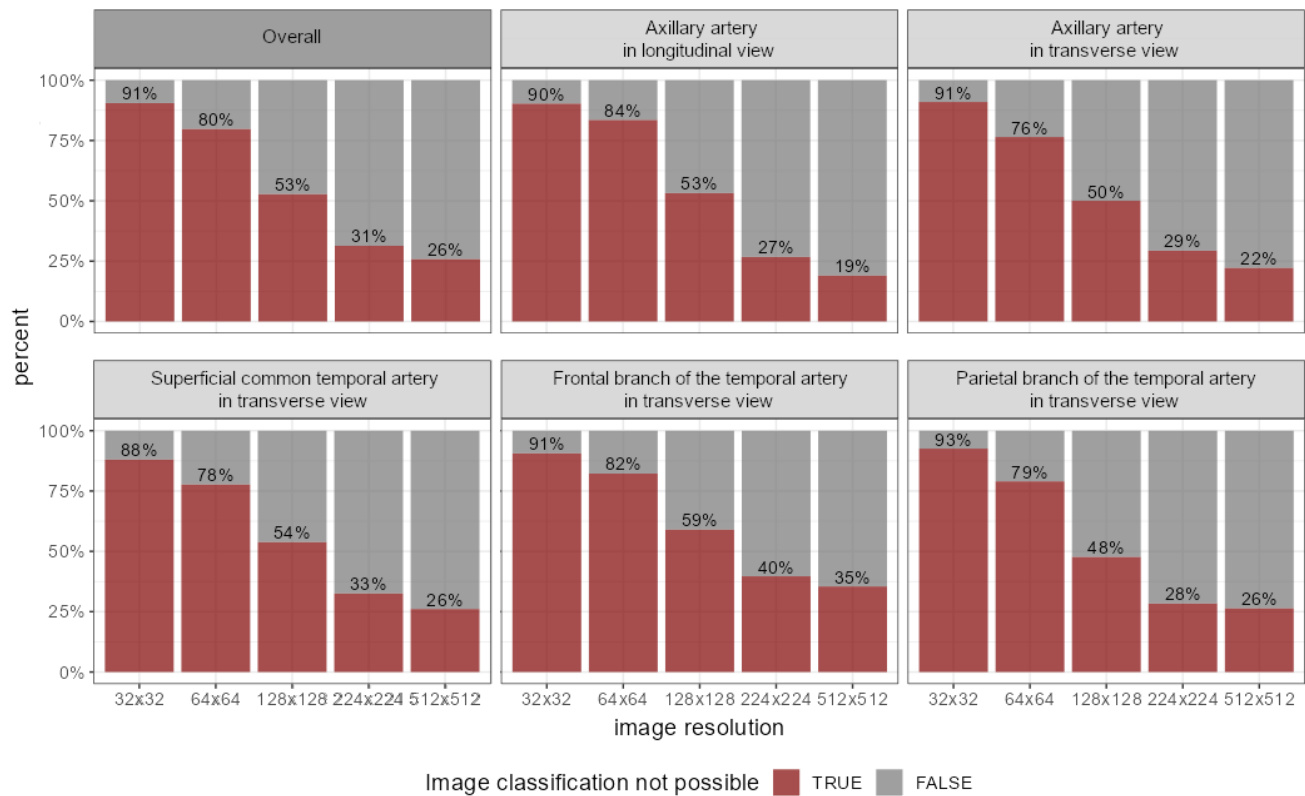


Figure 2. Frequency of the response option “Meaningful classification is not possible” being selected during vascular ultrasound image classification depending on the respective artery (and view) and image resolution. For all arteries, the frequency of this response option being selected decreased with increasing image resolution (constantly with every sequential stage of raised image resolution).

open.rsyd.dk

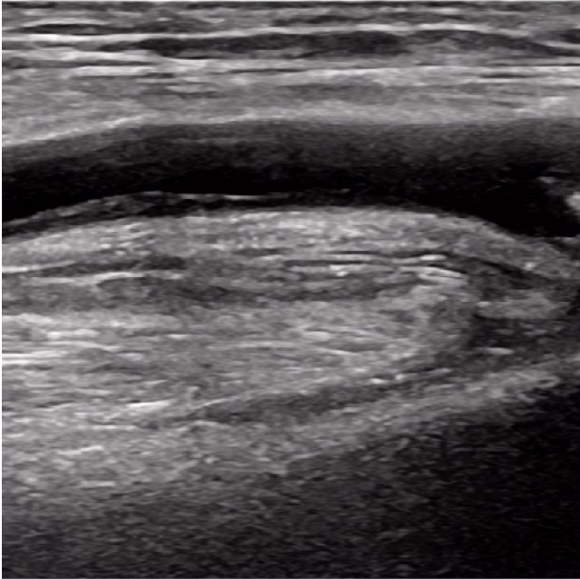
GCA images resolution project

INFO:
Throughout the following pages, you will be shown ultrasound images (one image per page). You will be kindly asked to classify each image into "GCA patient", "healthy individual (without GCA)", or "classification impossible". Each image and page depicts a separate case/patient.

Please note: Once you press "next" to be transferred to the next page, you will not be able to return to previous images and pages.

Please evaluate a series of 50 images of the axillary artery in longitudinal sections.

1



I classify this ultrasound image as belonging to a

☐ GCA patient
☐ Healthy individual
☒ Meaningful classification is not possible

If only one of the above-mentioned answers was accepted, I would pick:

☒ GCA patient
☐ Healthy individual

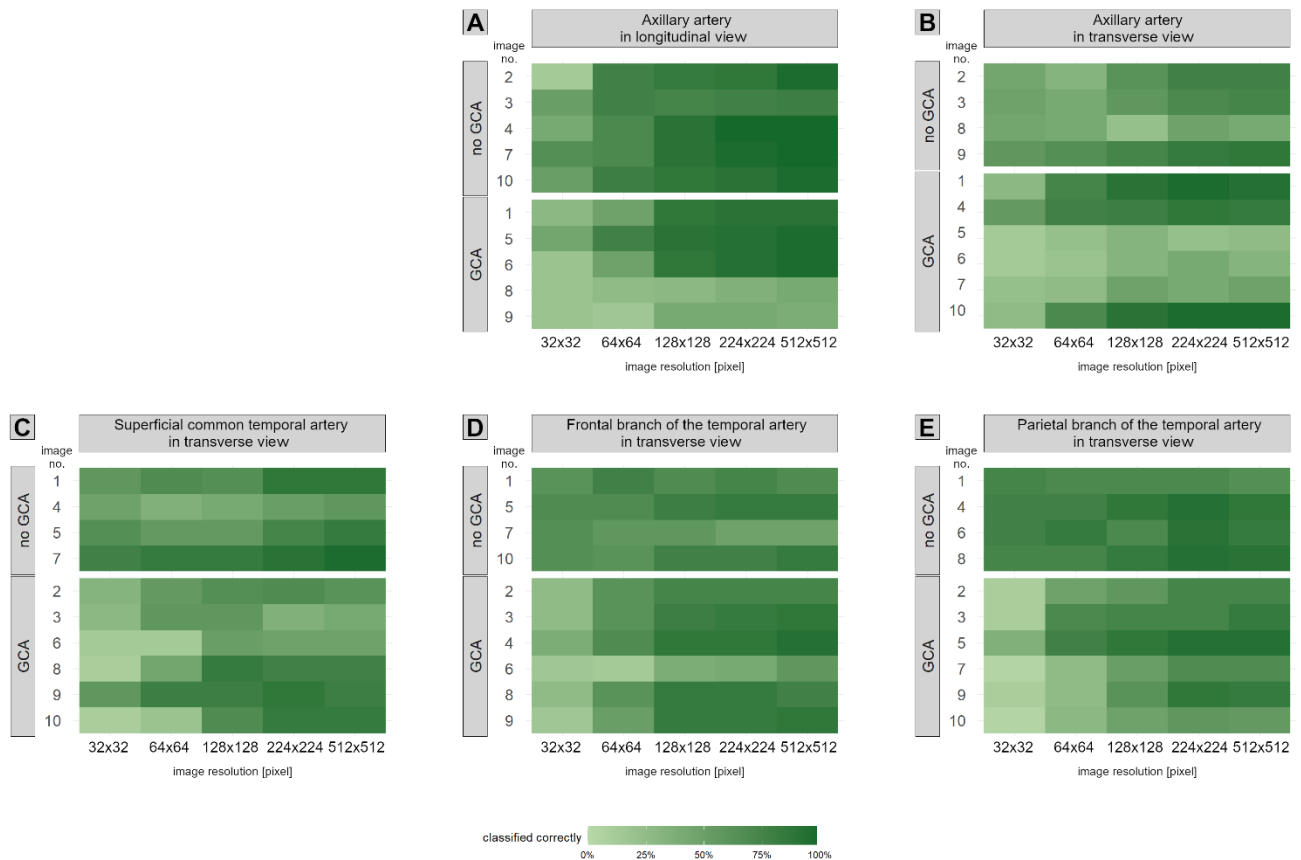
reset

Next Page >>

Powered by REDCap

Supplementary Figure 1. A screenshot of the web-based image classification exercise conducted via REDCap is shown. Participants reviewed vascular ultrasound images individually, with each image representing a separate patient case. The assessment was blinded, meaning participants had no access

to intima-media thickness measurements or clinical patient details. For each image, participants responded to the question, “*I classify this ultrasound image as belonging to a ...*” by selecting one of three options: (1) “GCA patient,” (2) “Healthy individual,” or (3) “Meaningful classification is not possible.” If the third option was chosen, a secondary dialog prompted participants to make a forced-choice decision between “*GCA patient*” and “*Healthy individual*” in response to the follow-up question, “*If only one of the above-mentioned answers was accepted, I would pick...*”



Supplementary Figure 2. Heatmap plot illustrating the frequency each individual image was classified correctly at different image resolutions, for: [A] axillary artery in longitudinal view, [B] axillary artery in transverse view, [C] superficial common temporal artery in transverse view, [D] frontal branch of the temporal artery in transverse view, [E] parietal branch of the temporal artery in transverse view (total number of correct classifications was applied, including both number of correct confident classifications plus number of correct forced classifications).

635
636
637
638
639
640
641

642 **Supplementary Material 1: Conception of the envisioned AI**

643 As ultrasound machines with transducers of sufficient quality are available in many hospitals
644 worldwide, but expertise in ultrasound diagnosis of GCA is not, it appears timely to leverage
645 technological advancements — such as artificial intelligence — to bridge existing gaps, facilitating
646 and accelerating the ultrasound diagnosis of GCA.

647 The envisioned AI-based GCA ultrasound image classification assistant will be integrated into a web-
648 based platform that (1) guides the acquisition of ultrasound images, (2) accepts the upload of acquired
649 image material, (3) performs AI-driven image classification, and (4) provides the user with a final
650 assessment.

651 The ultimate goal is to integrate all key elementary ultrasound lesions of GCA¹ that have been
652 established as reliable criteria in clinical practice²—including overall sonomorphology, like intima-
653 media thickening, the halo sign, and the compression sign—into the AI’s image evaluation and
654 classification process, ideally using dynamic ultrasound recordings. Future advancements may also
655 incorporate inflammatory activity levels and clinical features to further refine diagnostic accuracy.
656 However, the primary focus remains on supporting imaging-based GCA diagnostics.

657 To facilitate ultrasound acquisition even for less experienced examiners, a concise photo and video
658 guide will be provided, offering concise, step-by-step instructions for probe positioning and handling.
659 After image acquisition, the web-based platform will perform a preliminary quality check to ensure
660 that the uploaded material meets the minimum requirements for AI-based classification. Upon
661 approval, the AI-based classification will be conducted, and the system will generate a final assessment
662 regarding the presence of GCA-suspicious vascular changes. The user will receive a report via the web-
663 based platform, including example reference images that contributed to the AI’s classification decision,
664 thereby enhancing transparency and interpretability.

The development of the AI-based GCA ultrasound image classification assistant requires a multi-step approach involving both model training and validation. As an initial step, a simplified focus on overall OMERACT GCA definitions and intima-media thickening is proposed. Unlike the compression sign, which requires dynamic ultrasound acquisition, and the halo sign, which demands expertise in correctly adjusting power Doppler settings, these features can be assessed solely from static B-mode images. Consequently, the first phase of AI development will focus on detecting GCA-suspicious vasculitic changes from static B-mode images, which can be more easily acquired—even by less experienced users following proper guidance. The AI will initially evaluate one ultrasound image at a time, with the integration of multiple assessments from different images of the same patient occurring at a superordinate level. To prevent the AI from relying too heavily on potentially provided intima-media thickness measurements, which could shift its focus away from actual image analysis, the initial AI model will be trained exclusively on image data, independent of any pre-existing thickness values. To allow for high comparability, the methodology of the present study, which explored the minimum image resolution required for human expert classification of vascular ultrasound images in GCA and healthy subjects, was deliberately aligned with this setup. Determining the minimum resolution which is required for human expert classification, and thus also anticipated to be for AI-based classification, is a crucial step in the development of the envisaged AI-based GCA ultrasound image classification assistant, as it establishes which AI architectures may be immediately disqualified due to their inability to process input images at the necessary resolution.

In addition to the carefully considered selection of a suitable AI architecture that achieves the ideal balance between processing image data in the required high resolution and the lowest possible resource requirements, there are other basic requirements for the development of an AI model. Beyond computational and personal resources this includes sufficient amount of training material (images) for proper neural network training, marking one of the most crucial aspects for AI development success^{3,4}. While commonly used techniques in neural network training, such as data augmentation, help artificially increase the number of image variations⁵, it has become standard practice to base model training on a fully pre-classified dataset containing at least 1,000 to 2,000 images⁶. Once training and validation of the neural network model are completed, it will be followed up by clinical validation to proof reliability. After thorough testing has been applied by experts such as members of the Large Vessel Vasculitis subgroup of the OMERACT Ultrasound working group, access will be made publicly available.

696

697 References:

- 698 1. Chrysidis S, Duftner C, Dejaco C, et al. Definitions and reliability assessment of elementary
699 ultrasound lesions in giant cell arteritis: a study from the OMERACT Large Vessel Vasculitis
700 Ultrasound Working Group. 2018;4:e000598.
- 701 2. Schäfer VS, Chrysidis S, Dejaco C, et al. Assessing Vasculitis in Giant Cell Arteritis by
702 Ultrasound: Results of OMERACT Patient-based Reliability Exercises. *J Rheumatol*.
703 2018;45:1289–1295.
- 704 3. Sze V, Chen Y-H, Yang T-J, et al. Efficient processing of deep neural networks: A tutorial and
705 survey. *Proc IEEE*. 2017;105:2295–2329.
- 706 4. Pinto N, Cox DD, DiCarlo JJ. Why is real-world visual object recognition hard? *PLoS Comput*
707 *Biol*. 2008;4:e27.
- 708 5. Wu R, Yan S, Shan Y, et al. Deep image: Scaling up image recognition. *ArXiv Prepr*
709 *ArXiv150102876*;7.
- 710 6. Bressem KK, Vahldiek JL, Adams L, et al. Deep learning for detection of radiographic
711 sacroiliitis: achieving expert-level performance. *Arthritis Res Ther*. 2021;23:106.

712

713

Reply to editor and reviewers:

Manuscript Number: ARDS-D-24-00403

Exploring the Limit of Image Resolution for Human Expert Classification of Vascular Ultrasound Images in Giant Cell Arteritis and Healthy Subjects: The GCA-US-AI Project

Dear Dr. Bauer,

Thank you for submitting your manuscript to Annals of the Rheumatic Diseases.

I have completed my evaluation of your manuscript. Based on the reviewers' comments, I invite you to submit a new version of your manuscript following major revision after addressing the comments below. Please resubmit your revised manuscript by Mar 18, 2025.

When revising your manuscript, please consider all issues mentioned in the reviewers' comments carefully. Some of these comments may have to be addressed by performing new analyses or experiments, as outlined by the referees. When responding to the comments, please mention every change made in a clearly structured point-by-point response and provide suitable explanations for any comments not addressed. Please supply a tracked and a clean version of your paper upon resubmission. Also, please note that your revised submission will likely be re-reviewed.

To submit your revised manuscript, please log in as an author at <https://www.editorialmanager.com/ards/>, and navigate to the "Submissions Needing Revision" folder.

Research Elements (optional)

This journal encourages you to share research objects - including your raw data, methods, protocols, software, hardware and more – which support your original research article in a Research Elements journal. Research Elements are open access, multidisciplinary, peer-reviewed journals which make the objects associated with your research more discoverable, trustworthy and promote replicability and reproducibility. As open access journals, there may be an Article Publishing Charge if your paper is

746 accepted for publication. Find out more about the Research Elements journals at
747 [https://www.elsevier.com/authors/tools-and-resources/research-elements-](https://www.elsevier.com/authors/tools-and-resources/research-elements-journals?dgcid=ec_em_research_elements_email)
748 [journals?dgcid=ec_em_research_elements_email](https://www.elsevier.com/authors/tools-and-resources/research-elements-journals?dgcid=ec_em_research_elements_email).

749

750 Annals of the Rheumatic Diseases values your contribution and I look forward to receiving your
751 revised manuscript.

752

753 Kind regards,

754 Josef Smolen

755 Editor-in-Chief

756 Annals of the Rheumatic Diseases

757

758 Editor and Reviewer comments:

759

760

761 Dear Prof. Smolen, dear Reviewers,

762

763 First of all, thank you very much for your thorough review, valuable input and great advice. The time
764 and effort put in strongly comes through and is highly appreciated, as is the additional input!

765

766 In order to do justice to this commitment, we will be more than happy to address all comments point
767 by point below.

768

769

770

771 **Reviewer 1:**

772 Bauer et al. conducted this study to investigate the minimum image resolution required for human
773 experts to reliably classify ultrasound images of arteries as being affected by GCA or not. The authors
774 convened an expert panel who evaluated images at different levels of resolution - with results laying
775 the groundwork for eventual development of AI tools. The study is highly innovative, was conducted
776 by experts in the field and the data is presented very clearly. The international collaboration is a notable
777 strength of the study. Some aspects deserve further clarification:

1. The images were derived from subjects with GCA, and from healthy controls. In clinical practice, patients may have GCA mimics such as atherosclerosis, fibromuscular dysplasia and other vasculopathies. The study would be strengthened by the inclusion of such mimics as perhaps a different image resolution would be required to accurately classify such patients vs GCA.

Our Response -> Thank you very much for this insightful comment which provided us highly appreciated input to improve parts of the manuscript. We fully acknowledge the importance of differentiating GCA from other vasculopathies such as atherosclerosis, fibromuscular dysplasia, and other mimics, although it must be fairly admitted that distinguishing some of these differential diagnoses – other than atherosclerosis – from GCA is not feasible via ultrasound. Important contribution to this challenge has been made by previous research that demonstrated particularly sonography of the temporal and axillary arteries reduces bias of these differential diagnoses in contrast to other vessels that are more predisposed to present with GCA mimics. Fibromuscular dysplasia, e.g., is commonly encountered in the renal, carotid, and vertebral arteries whereas it rarely affects temporal and axillary arteries. While atherosclerotic changes can occur in the axillary arteries, they are typically distinguishable by ultrasound imaging based on their characteristic morphology, allowing experienced readers to differentiate them from GCA-related changes reliably (<https://doi.org/10.1136/rmdopen-2017-000598>). In contrast, the temporal arteries are muscular arteries and rarely affected by atherosclerosis, as atherosclerosis primarily affects large elastic arteries such as the aorta, coronary arteries, and carotid arteries. Internationally established guidelines such as the EULAR recommendations for the use of imaging in large vessel vasculitis in clinical practice (<https://doi.org/10.1136/annrheumdis-2017-212649>) reflect the high suitability of temporal and axillary artery ultrasound for distinguishing these conditions.

Given these considerations, our study followed these recommendations and focused on both the temporal and axillary arteries as scientifically validated and clinically well-established imaging sites for GCA assessment.

That said, we fully agree with your suggestion that future research should explore whether higher image resolution is needed to differentiate GCA from specific mimics, particularly in borderline cases where classification might be more challenging. We added a new paragraph based on your comment to the manuscript's discussion section.

2. The authors indicate that some patients were diagnosed by biopsy, and some by ultrasound. What was the proportion of patients diagnosed by biopsy? Did any of the patients have a change in clinical diagnosis during the follow-up period (to a non-GCA diagnosis)?

Our Response -> Thank you for your valuable question. While we confirmed that each GCA patient of which ultrasound images were used had confirmed diagnosis by either temporal artery biopsy or

imaging, including ultrasound, we did not report specific proportions of patients diagnosed by biopsy versus ultrasound, as the classification of patients in our study was based on adherence to the classification criteria (extended 2016 ACR or 2022 ACR/EULAR classification criteria for GCA, depending on the date of diagnosis) rather than the initial diagnostic approach. Going beyond this singular form of investigation, these established classification criteria consider various different modalities for correct classification, including clinical evaluation, imaging, and histopathology, if used at that time.

Regarding the follow-up period, none of the patients had a change in clinical diagnosis from GCA to a non-GCA diagnosis. This confirms the stability of the initial classification over time.

Inspired by your input we revised section '3.2 Vascular ultrasound image source and selection' for further clarification.

Reviewer 2:

Important paper on US in GCA in which 31 international experts in the field participated.

In the introduction it is stated that the arrival of modern handheld ultrasound devices, characterized by high affordability (<US\$3000), could further expand access to point-of-care diagnostic imaging, particularly in resource-limited settings. This advancement also has the potential to create new educational opportunities for training both specialists and non-specialists in scanning and interpreting arterial findings in the context of GCA.

A draw back could be the low quality of the images of the hand held machines this should be mentioned.

Our Response -> Thank you very much for this comment which we agree with! We appreciated your suggestion and revised the manuscript accordingly.

Also in the following sentence: GCA ultrasound has proven its excellent reliability among experts with adequate ultrasound equipment. The definition of adequate should be specified.

Our Response -> We have been grateful for this comment and have added further specification of "adequate ultrasound equipment" to the manuscript under section "2 Introduction".

As also reflected by 2023 published "Quality standards for the care of people with giant cell arteritis in secondary care" (doi.org/10.1093/rheumatology/kead025), adequate ultrasound equipment is commonly defined as an ultrasound machine with a linear ≥ 10 MHz probe for axillary artery imaging and a ≥ 18 MHz linear probe for cranial artery examination.

However, no attempt has yet been made to apply AI in GCA to classify B-mode ultrasound images as abnormal or normal. B mode ultrasound should be explained.

Our Response -> Thank you for this suggestion. We have now included a brief explanation of B-mode ultrasound in the introduction for better clarity.

Also in the introduction a lot of background information is provided on AI which can partly be removed the same goes for the discussion the article should focus on how to improve the CDUS as diagnostic tool for GCA and focus and the minimal requirements for trained sonographers to be able to rule in or out GCA.

Our Response -> We sincerely appreciate your valuable feedback. While we fully acknowledge the importance of optimizing CDUS as a diagnostic tool for GCA and defining the minimal training requirements for sonographers, these aspects fall beyond the primary scope of our study. Our focus was specifically on determining the minimum image resolution required for human expert classification of B-mode ultrasound images, a critical step in developing AI-based tools for GCA diagnosis.

The key consideration here is that, for the development of the envisioned AI-based GCA ultrasound image classification assistant, it is crucial to determine how far ultrasound image resolution can be reduced - to make it feedable to as many AI architectures as possible - while still maintaining classifiability according to human experts, serving as a reference.

That said, we greatly appreciate your comment, which, along with other reviewer suggestions, has guided us in elaborating Supplementary Material 1 "Conception of the envisioned AI". This addition clarifies the underlying rationale of our study and the broader implications for AI development in the field of GCA imaging.

Methods

Important to say something about the minimal requirements of the equipment; Each image was derived from an individual patient or healthy subject. Images were obtained from different ultrasound machine brands (Esaote, GE and Philips) using 3-12 (Philips Epiq), 4-12 (GE LOGIQ e), 6-15 (GE Logiq S8), 5-18 (Philips Epiq), 6-18 (Esaote MyLab 8), 8-18 (GE Logiq E9) and 6-24 MHz (GE Logiq E10) linear transducers for the investigation of axillary arteries and 5-18 (Philips Epiq), 8-18 (GE Logiq S8, GE Logiq E9) and 6-24 MHz (GE Logiq E10) hockey stick probes for the investigation of cranial arteries. Image acquisition at each site was performed according to a previously published scanning protocol. Is the equipment state of the art and could the variation in equipment influence the data? Also explain why the single static B mode images are the preferred images. Lastly since a lower image resolution had a higher chance that the image was marked as "meaningful classification is not possible". compared

to a resolution of 512x512 pixels, underlines that there minimal requirements for the US machine are needed. Despite the fact that maybe experts can also interpret lower quality images. Since you need to rule in or out GCA the test should have a high sensitivity and specificity.

Our Response -> Thank you for your thoughtful input. We fully agree that minimal requirements for ultrasound machines exist, as outlined in the 2023 publication *Quality standards for the care of people with giant cell arteritis in secondary care* (doi.org/10.1093/rheumatology/kead025). We have now included this reference and emphasized the recommended use of ≥ 10 MHz linear probes for axillary artery imaging and ≥ 18 MHz for temporal artery examination. The ultrasound equipment used in this study was state-of-the-art, with intentional variation to enhance the generalizability of our classification exercise. Additionally, based on the valuable input of this and other reviewer comments, we have addressed the rationale for using single static B-mode images in a newly added “Supplementary Material 1: Conception of the envisioned AI”. Addressing the last section of your train of thought that we gratefully received, it might be noteworthy to indicate a subtle distinction: While our findings may indeed allow for broader implications regarding image resolution requirements in ultrasound diagnostics, the primary focus of our study was not to determine the optimum image resolution ultrasound machines need to provide on their display so that readers can sufficiently work with this ultrasound machine. Instead, our study’s main goal was to explore how far ultrasound image resolution can be decreased to make it feedable to as many AI architectures as possible while still remaining classifiability according to human experts serving as a reference. That being said, we fully support your view that further research into the minimum requirements for ultrasound machines, as well as the refinement of standards and recommendations, is essential to enhance the quality and consistency of ultrasound-based GCA diagnostics.

Discussion

It should be mentioned that high end US machines are needed for good quality data.

Our Response -> Thank you very much for this comment which we agree with and which we feel to have addressed simultaneously by enriching the manuscript with details on what defines adequate ultrasound equipment, a reference to the 2023 published “Quality standards for the care of people with giant cell arteritis in secondary care” that supports your statement, and additional information regarding the limitations in the quality of portable ultrasound devices that currently still exist and will need to be overcome in the future, according to your previous input.

Discuss why only one single static B-mode image per patient without any additional IMT measurement were used.

Our Response -> We appreciated your impulse, which we addressed by adding in-depth information on our vision and conception of the envisioned AI-based GCA ultrasound image classification assistant for further clarification. Due to the length restrictions of the main text this was handled as a new supplementary material, that is now also referred to by the manuscript (see Supplementary Material 1: Conception of the envisioned AI).

Presently the importance of the more objective IMT measurement is recognized and advised. Since the idea is to develop a training model for AI it would be good to build a more complex model including the HALO count since this probably will increase the sensitivity and specificity of the data.

Our Response -> We appreciate this forward-thinking suggestion and agree that incorporating HALO count in future AI models could enhance sensitivity and specificity. In order to properly address your comment as well as a few other reviewer comments, we decided to extend on our vision and conception of the envisioned AI-based GCA ultrasound image classification assistant by adding further in-depth information. Due to the length restrictions of the main text this was accomplished by the addition of a new supplementary material, that is now also referred to by the manuscript (see Supplementary Material 1: Conception of the envisioned AI).

Also discuss the minimal requirements for deep neural network training

Our Response -> Thank you very much for this suggestion, which we addressed – together with a few other reviewer comments - by the addition of Supplementary Material 1.

Minor points

Introduction: The arrival of modern handheld ultrasound devices, characterized by high affordability (<US\$3000), would adept this in Euro's.

Our Response -> Thank you very much for this suggestion, which we fully agree with, especially considering the European focus of the target journal. To account for currency fluctuations while maintaining readability, we have now provided an approximate conversion rate and specified the price as <€3000 in the manuscript.

Last sentence required for human experts to reliably classify ultrasound images of arteries commonly affected by GCA for the presence or absence of the disease. Change into that you want to reliably rule in or out vasculitis

Our Response -> Thank you very much for your thoughtful suggestion regarding the wording. While we appreciate your proposed wording with regards to its ease of comprehensibility, we acknowledge

that the originally chosen phrasing, also including the term "classify" in particular, precisely aligns with the standard scientific terminology used in both fields GCA ultrasound diagnostics and AI image classification within this specific application context. To maintain consistency with established terminology, we would suggest retaining the original phrasing, provided your consent.

Reviewer 3:

This study addresses an important challenge in clinical practice: the early diagnosis of GCA and the interobserver reliability of temporal artery ultrasound. Developing tools such as AI-based software could improve accuracy and precision in diagnosis, especially considering recent advancements like the OGUS score for evaluating the Halo sign across vascular regions. Thus, the study could represent an interesting foundation for further work on developing AI models. However, I am afraid that the premise that AI application could help centers where US expertise is not available is not in line with the results or the aims of the study. In fact, the problem of image resolution and recognition is absolutely secondary compared to the correct imaging acquisition.

Our Response -> Thank you for sharing this important thought process with us! In order to properly address your comment as well as a few other reviewer comments, we decided to extend on our vision and conception of the envisioned AI-based GCA ultrasound image classification assistant by adding further in-depth information. Due to the length restrictions of the main text this was accomplished by the addition of a new supplementary material, that is now also referred to by the manuscript (see Supplementary Material 1: Conception of the envisioned AI).

One strength of the study is the methodologically correct acquisition of images (following a standardized protocol), although different machines and probes were used for image acquisition. However, a major limitation, as highlighted by the authors themselves, is the static B-mode imaging used in this study. The absence of essential elements like intima-media thickness measurement, Halo sign, compression sign, or clinical information, significantly deviates from the dynamic nature of ultrasound and its natural interpretation in clinical practice. This is likely a primary reason for the high rate of initially "non-classifiable" images.

Besides that, I would suggest the authors to address a few important issues in order to make their conclusions more solid and clear:

1) One the most relevant bias is that even people who contributed images themselves were involved in image interpretation. Why? Would it not have been sufficient only to use classification from the other experts of the group? This is actually not even a confounding factor, it should be considered a study bias, as statistical analyses might not be sufficient to address this issue.

Our Response -> First of all, let me thank you for this important and thoughtful remark. I suppose additional insight into the study conduct and thought process behind might help to address your concerns. To start with, it is relevant to mention that the entire image selection, processing and preparation was performed by CJB, who did not participate as an expert in image classification in this study. Also, he was not specifically provided with a handful of pre-selected images for this study but gained access to entire libraries out of which only a fraction of images was randomly selected. These dated back months up to years, reducing the likelihood that creators could have specifically recognized the individual images. Still, it was vital for us to challenge our own argumentation: To further ensure robustness, we performed sensitivity analyses for all metrics reported under sections "4.2 Image classification results", "4.3 Sensitivity and specificity" and "4.4 Factors impacting the classification decision" after the exclusion of all five image contributors. While slight numerical differences could be observed as expected, exclusion of these particular study participants did not lead to a significant change in the results. As indicated by your comment and in sincere appreciation of your valuable input we revised the limitations section of our manuscript for further transparency. Moreover, even though the results of this further investigation comprise a large dataset, we would be happy to display this data in detail as an additional supplement in case you found this beneficial and advocated for this approach.

2) Although the study refers to the use of low-cost machines to improve accessibility, the equipment used for image acquisition in this study does not align with that claim

Our Response -> We agree with this comment and acknowledge that the manuscript provides an outlook to the future at the point of this reference. Of course, this is naturally in line with the motivation behind this project and scientific efforts in general to outgrow the present and shape the future of patient care. However, that being said, your comment inspired us to revise the manuscript for further clarification.

3) It is unclear how significant confounding factors (e.g., image resolution per examiner, level of experience, and so on) were addressed. Clearer details on the models used to account for these factors would be helpful: in particular, lines 220-224, the outcome of the logistic mixed effect model is not even explicitly stated: what is "improvement or deterioration in image classification performance"? How was it defined?

1017 **Our Response ->** Thank you for highlighting this point. In order to clarify the used logistic mixed
1018 effects models and to improve interpretation of the results, we have now revised the manuscript by
1019 adding further information to the “Materials and Methods” section.

1020

1021 4) Could you provide results of this logistic mixed-effects model? How was image resolution
1022 considered? As a Dummy variable? What was the reference category? Regarding the model, it remains
1023 unclear the fact that "The number of years experts had been interpreting ultrasound images of giant
1024 cell arteritis did not have a significant impact on the selection of the response category "Meaningful
1025 classification is not possible." and "The estimated number of patients experts had seen and evaluated
1026 for suspected giant cell arteritis over the course of their careers did not significantly affect the
1027 evaluation." are valid for all resolution categories or not, not knowing the exact selection and
1028 definitions of the main independent variables and covariates.

1029 **Our Response ->** Thank you very much for this important remark that led us to a revision of the
1030 original manuscript content for further classification. As stated by our latest addition to the “Materials
1031 and Methods” section, the image resolution was included as a categorical independent variable. The
1032 image resolution (in pixel) used as the reference category, situationally depended on the intended
1033 direction of the analysis but was applied as described in the respective results. With regards to the
1034 logistic mixed-effects model you specifically refer to by this comment, a resolution of 512x512 pixel
1035 was used as the reference category.

1036 Further you refer to our statement “The number of years experts had already been reading GCA
1037 ultrasound images had no significant effect on the selection of the response category “Meaningful
1038 classification is not possible” (OR= 1.02, p=0.725). Neither did the number of patients that experts
1039 estimated to have seen and evaluated for suspected GCA throughout their career (OR= 1.00,
1040 p=0.422).” under section “4.4 Factors impacting the classification decision”. The included statements
1041 are valid for all resolution categories. For further reference, we are happy to provide you with in-depth
1042 insight into the results of this logistic mixed-effects model below:

1043

Effect of the number of years experts had already been reading GCA ultrasound images on the selection of the response category “Meaningful classification is not possible”

The selection status of the response category “Meaningful classification is not possible” (classification 0 = possible / 1 = not possible) is the dependent variable. Artery type, image resolution, and number of years reading GCA ultrasound images are included as independent variables (fixed intercept). Since there are repeated measurements for both the 31 readers and the 10 images at different resolutions, these are included as random effects in the model (random intercept).

	<i>Model <u>not</u> adjusting for differences across arteries</i>			<i>Model <u>adjusting</u> for differences across arteries</i>		
	Selection of the response category “Meaningful classification is not possible”			Selection of the response category “Meaningful classification is not possible”		
<i>Predictors</i>	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>
(Intercept)	0.15	0.04 – 0.61	0.008	0.13	0.03 – 0.54	0.005
Number of years reading GCA ultrasound images	1.02	0.91 – 1.14	0.726	1.02	0.91 – 1.14	0.725
Image resolution [224x224]	1.42	1.19 – 1.70	<0.001	1.43	1.19 – 1.71	<0.001
Image resolution [128x128]	4.96	4.14 – 5.95	<0.001	5.05	4.21 – 6.06	<0.001
Image resolution [64x64]	39.69	31.35 – 50.26	<0.001	40.97	32.31 – 51.95	<0.001
Image resolution [32x32]	190.36	138.66 – 261.35	<0.001	198.84	144.51 – 273.59	<0.001
Artery [Axillary artery - transverse view]				0.94	0.77 – 1.15	0.538
Artery [Frontal branch - transverse view]				1.75	1.43 – 2.14	<0.001
Artery [Parietal branch - transverse view]				1.02	0.84 – 1.25	0.837
Artery [Superficial common temporal artery - transverse view]				1.09	0.89 – 1.34	0.383
Random Effects						
σ^2	3.29			3.29		
τ_{00} (patients)	3.80			3.87		
τ_{00} (images per artery and resolution)	0.11			0.12		
ICC	0.54			0.55		
n (number of patients)	31			31		
n (number of images per artery and resolution)	10			10		
Observations	7750			7750		
Marginal R ² / Conditional R ²	0.359 / 0.707			0.363 / 0.712		

1044
1045

Effect of the number of patients with suspected GCA they have evaluated throughout their career on the selection of the response category "Meaningful classification is not possible"

The selection status of the response category "Meaningful classification is not possible" (classification 0 = possible / 1 = not possible) is the dependent variable. Artery type, image resolution, and number of patients evaluated are included as independent variables (fixed intercept). Since there are repeated measurements for both the 31 readers and the 10 images at different resolutions, these are included as random effects in the model (random intercept).

<i>Predictors</i>	<i>Model not adjusting for differences across arteries</i>			<i>Model adjusting for differences across arteries</i>		
	Selection of the response category "Meaningful classification is not possible"			Selection of the response category "Meaningful classification is not possible"		
	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>
(Intercept)	0.16	0.07 – 0.37	<0.001	0.14	0.06 – 0.32	<0.001
Number of patients evaluated	1.00	1.00 – 1.00	0.422	1.00	1.00 – 1.00	0.422
Image resolution [224x224]	1.42	1.19 – 1.69	<0.001	1.42	1.19 – 1.70	<0.001
Image resolution [128x128]	4.91	4.09 – 5.88	<0.001	4.99	4.16 – 5.98	<0.001
Image resolution [64x64]	38.77	30.65 – 49.03	<0.001	39.99	31.57 – 50.66	<0.001
Image resolution [32x32]	183.89	134.21 – 251.96	<0.001	191.94	139.79 – 263.55	<0.001
Artery [Axillary artery - transverse view]				0.94	0.77 – 1.15	0.540
Artery [Frontal branch - transverse view]				1.74	1.42 – 2.13	<0.001
Artery [Parietal branch - transverse view]				1.02	0.84 – 1.25	0.838
Artery [Superficial common temporal artery - transverse view]				1.09	0.89 – 1.34	0.386
Random Effects						
σ^2	3.29			3.29		
τ_{00} (patients)	3.74			3.80		
τ_{00} (images per artery and resolution)	0.11			0.12		
ICC	0.54			0.54		
n (number of patients)	31			31		
n (number of images per artery and resolution)	10			10		
Observations	7750			7750		
Marginal R^2 / Conditional R^2	0.362 / 0.706			0.366 / 0.711		

5) What is the meaning of "forced classification"? why was this necessary?

Our Response -> We appreciate this request for clarification. It is in the nature of the investigated levels of image resolution in this study to give rise to the expectation that expert uncertainty in classification might occur, at least to some extent, particularly at lower image resolution levels. In order to obtain the most informative results in this study, we opted for a two-step answering approach.

As a first step, participants could either provide a definitive classification of an image (as belonging to a “GCA patient” or “Healthy individual”) or indicate that classification was not possible. This first answer was referred to as ‘confident classification’. However, to further understand the participants’ decision-making process, we also asked them to provide a definitive classification choice, as a second step (referred to as ‘forced classification’), in cases where they had initially answered that a meaningful classification is impossible. This allowed us to gain additional insights into the underlying classification tendencies and the impact of image resolution on expert confidence, as reported by the manuscript.

6) It is unclear if all examiners followed the same protocol regarding forced classifications. For instance, after an initial response of "classification not possible," REDCap allowed some participants to skip the forced classification and proceed to the next image. This discrepancy raises concerns about potential inconsistencies in data collection. In particular, the rationale for forcing classification in some cases but not in others, is not clearly explained. This inconsistency could introduce bias, especially for ambiguous cases.

Our Response -> Thank you very much for your thoughtful observation. To ensure consistency in the classification process, we deliberately configured the REDCap platform to display only one ultrasound image at a time, prevent participants from returning to previous images, and require completion of each classification before proceeding. This approach was intended to minimize potential bias by preventing retrospective adjustments or influences on subsequent classifications. However, as we later discovered, a technical error in REDCap unintentionally allowed participants to proceed to the next image without completing the forced classification follow-up question whenever they had opted for “meaningful classification is not possible” in first instance. Importantly, this issue affected all participants equally, and the number of missing forced classifications remained minimal. In great appreciation for your input, for clarification and full transparency, we revised the methods section you refer to with this comment accordingly and also added this aspect to the limitations of this study.

7) The removal of missing classifications could introduce bias. The authors should discuss how missing data were handled and whether alternative approaches were considered.

Our Response -> Thank you very much for this important remark. We fully acknowledge that missing classifications could introduce bias and carefully considered various approaches to address this issue. Repeating the classification exercise with the same participants would have introduced substantial bias (due to memorability and recognizability), and as we had already asked the world-leading experts in this field to participate in the study assembling a similarly high-caliber international expert group was not a feasible alternative. We also discussed whether to uniformly categorize all missing classifications as incorrect, but this would have led to a considerably greater data bias, ultimately

making this approach unsuitable. Fortunately, as noted in our response to Comment 6, the missing classifications only affected the "forced classifications" and not the "confident classifications." To ensure full transparency, Table 1 provides a comprehensive breakdown of classification results across all arteries, views, and image resolutions, allowing readers to assess both confident and forced classification outcomes in detail. In acknowledgment of your valuable remark and our considerations, we have revised the limitations part in the discussion section accordingly.

8) While the highest specificity was achieved in frequently affected sites like the axillary arteries, these are less clinically relevant for the primary challenge of GCA diagnosis—preventing vision loss due to cranial involvement. The authors should discuss the implications of this finding for clinical practice.

Our Response -> Thank you very much for highlighting this important aspect of the overall debate in GCA care! We fully agree that vision loss is one of the most feared complications of GCA and represents a central concern in patient care. A natural approach to assessing this risk most directly would involve the evaluation of the central retinal artery or the optic nerve, as described in two of our most recent publications (References: doi.org/10.1093/rheumatology/keae287 and <https://doi.org/10.1007/s00296-025-05800-y>). However, this technique has yet to be widely implemented in clinical practice, requiring even greater expertise than conventional vascular ultrasound as well as dedicated high-end ultrasound equipment. Instead, for the global detection of giant cell arteritis, vascular ultrasound of the axillary and temporal arteries remains the recognized first-line diagnostic modality (References: doi.org/10.1136/annrheumdis-2017-212649, and doi.org/10.1093/rheumatology/kex461). For this reason - following current standard of care - we also emphasize the successful detection of disease-specific abnormalities in these two vessels in this manuscript, as we state: *"Our reported sensitivity of 0.833 and specificity of 0.915 for 224x224 pixel longitudinal view images of the axillary artery, as well as 0.734 and 0.855, respectively, for 224x224 pixel transverse view images of the superficial common temporal artery, align with data from a systematic review and meta-analysis investigating the diagnostic performance of temporal artery ultrasound (considering full-resolution images) for the diagnosis of GCA³⁷."* It is noteworthy that the classification of the ultrasound images of the temporal arteries (as representatives of significantly smaller and cranial vessels) succeeds with almost comparable specificity as in the axillary arteries.

9) The potential impact of hardware variability (e.g., monitor resolution, size, and quality) on image interpretation is not discussed but could be significant. If the involved participants were not aware that the objective of the exercise was finding the minimum image resolution required for human experts to

reliably classify ultrasound images of arteries commonly affected by GCA, they could have modified the resolution of the image, or magnifying it, in an attempt to better classify it.

Our Response -> Thank you very much for bringing up this train of thought that many may have in their minds, but which can be sufficiently refuted based on our study design. Starting point of our study were native image extracts in size 512x512 pixels. These images were subsequently downsized to all study-specific resolution variants, namely 32x32, 64x64, 128x128 and 224x224 pixels. Natively, particularly the low-resolution image variants would appear very small on a computer monitor. In order to ensure a reasonable and consistent display size on-screen throughout the web-based image classification exercise, we subsequently upscaled all images of all resolution variants to a uniform size of 1024x1024 pixels (without interpolation). 1024x1024 pixels represents an image size that covers a large portion of standard computer monitors without fully occupying the screen, allowing for appropriate viewing conditions. However, it is crucial to emphasize that this upscaling step does not alter the image content. A 32x32 pixels image that is subsequently upscaled to 1024x1024 pixels does gain additional details or informative pixels, but just displays any previous pixel larger in size. The same effect applies if study participants had attempted to manually modify image resolution, magnify the image or use different monitor resolutions, sizes and qualities. As a result, these hardware-related factors have an extremely subordinate impact on the study results and do not compromise the validity of our findings.

Minor observations:

1. The study describes an increase in classification accuracy with higher resolution up to 224x224 pixels, but the lack of significant improvement beyond this resolution is not well-explained. In particular, the "improvement in classification" threshold beyond which increasing resolution would have been useless/inefficient was not pre-determined.

Our Response -> Thank you very much for pointing this out. We acknowledge that defining a pre-determined threshold for classification improvement, beyond which further increases in resolution would be considered useless/inefficient, would have been an alternative approach. For example, this could have been an additional classification performance increase of less than 5%, as we observed it in fact for all vessels and sections in forced classification performance when increasing the image resolution from 224x224 pixels to 512x512 pixels. However, since no prior data existed regarding the minimum image resolution required for expert classification of vascular ultrasound images in GCA, we opted for an empirical approach. Our results showed a plateauing of classification performance at 224x224 pixels, with no significant improvement beyond this resolution. This aligns with our statistical analysis, where the odds of correct classification did not significantly increase from 224x224 pixels to 512x512 pixels. Evidently, a sufficiently high image resolution is achieved at 224x224 pixels, which

enables experts to clearly recognize and assign vascular abnormalities, explaining the lack of significant improvement beyond this resolution. In contrast, a higher image resolution with more pixels does not appear to offer any relevant added value.

2. The study does not address whether patient characteristics (e.g., comorbidities such as widespread atherosclerosis) could have influenced classification outcomes.

Our Response -> Thank you for this insightful comment which we addressed by adding a new paragraph to the discussion and limitations section. We fully acknowledge that patient-specific factors, such as widespread atherosclerosis, in theory bear the potential to impact ultrasound image interpretation. While atherosclerosis is uncommon in temporal arteries, atherosclerotic changes can occur in the axillary arteries, though are typically distinguishable by ultrasound imaging based on their characteristic morphology, allowing experienced readers to differentiate them from GCA-related changes reliably (<https://doi.org/10.1136/rmdopen-2017-000598>). Moreover, important contribution to the challenge of differentiating GCA from other vasculopathies has been made by previous research that demonstrated particularly sonography of the temporal and axillary arteries reduces bias of these differential diagnoses in contrast to other vessels that are more predisposed to present with GCA mimics. Fibromuscular dysplasia, e.g., is commonly encountered in the renal, carotid, and vertebral arteries whereas it rarely affects temporal and axillary arteries. Internationally established guidelines such as the EULAR recommendations for the use of imaging in large vessel vasculitis in clinical practice (<https://doi.org/10.1136/annrheumdis-2017-212649>) reflect the high suitability of temporal and axillary artery ultrasound for distinguishing these conditions.

Given these considerations, our study followed these recommendations and focused on both the temporal and axillary arteries as scientifically validated and clinically well-established imaging sites for GCA assessment. Furthermore, we decided neither to actively include nor exclude ultrasound images containing aspects of atherosclerosis. Instead, our dataset represents a random sampling from real-world clinical cases. Since our study focused on determining the minimum image resolution required for classification (GCA-indicating findings present or not) rather than evaluating differential diagnoses, we did not stratify cases based on comorbidities.

That said, we have gratefully captured your train of thought pointing towards future research that should explore whether higher image resolution is needed to differentiate GCA from these specific mimics, particularly in borderline cases where classification might be more challenging. This inspired us to add a new paragraph based on your comment to the manuscript's discussion section.

3. Given the significant variation in participants' average experience and the number of monthly scans performed, the study should clarify what was the threshold to define a reader as "expert."

1192 Naturally, these were all members of the Large Vessel Vasculitis subgroup of the Outcome Measures
1193 in Rheumatology (OMERACT) Ultrasound group, but was there a threshold to include members in
1194 this specific exercise?

1195 **Our Response** -> Thank you very much for highlighting this important aspect. We fully acknowledge
1196 that the definition of an “expert” in giant cell arteritis (GCA) ultrasound is not yet standardized. While
1197 work from Georgina Ducker and Chetan B. Mukhtyar (Reference:
1198 doi.org/10.1093/rheumatology/keac591) has demonstrated the feasibility of specialist nurses
1199 performing GCA ultrasound assessments and becoming a reader in clinical routine after structured
1200 training and a minimum of 15 supervised examinations over five months, we consider the designation
1201 of "expert" in GCA ultrasound to require a more extensive level of experience.

1202 For this study, we established clear inclusion criteria to define an expert. In the absence of a formal
1203 ultrasonography certification (Reference: doi.org/10.1093/rheumatology/kead025), we considered the
1204 following as prerequisites: (1) being a board-certified rheumatologist, (2) having at least four years of
1205 experience in interpreting GCA ultrasound images, and (3) having evaluated a minimum of 100
1206 patients with suspected GCA throughout the career. Additionally, all participants were members of the
1207 Large Vessel Vasculitis subgroup of the Outcome Measures in Rheumatology (OMERACT)
1208 Ultrasound Group and had engaged in scientific work related to GCA.

1209 Following your suggestion of clarification we revised section “3.1 Study design” of the manuscript
1210 accordingly.

1211

1212